The Statistical Estimation and Adjustment Process using a PPSWR sampling design in the Vegetation Resource Inventory- Summary.[1]

Carl James Schwarz
Department of Statistics and Actuarial Science
Simon Fraser University
Burnaby, BC V5A 1S6

10 February 2001

This document is a highly condensed summary of Schwarz (2000) [*The Statistical Estimation and Adjustment Process using a PPSWR sampling design in the Vegetation Resource Inventory*] produced under the same contract. The latter should be consulted for details and examples of using the proposed design.

---

[1] This report was prepared for the Ministry of Forests, Government of British Columbia, Canada as part of contract 52MFR00015VS.

1. Introduction

The Vegetation Resource Inventory (VRI) is a provincial survey that collects data on the type, amount, and location of vegetation in British Columbia. The survey is carried out in inventory units, which may be as small as a watershed or as large as a Forest District. The survey process consists of two phases. In Phase I, current information for the entire inventory unit is obtained. Generally, aerial photographs are taken of the entire inventory unit. These are used to divide the inventory unit into homogenous areas called polygons and to provide photo-interpreted estimates for every polygon in the inventory unit. However, in some cases current information is obtained by updating a past inventory. In Phase II a sample of polygons are selected for taking ground measurements. These ground measurements are then used to adjust the estimated inventory totals as well as the polygon estimates from Phase I accounting for the observed data from the ground points.

The Ministry of Forests decided to change the sampling methodology for the Vegetation Resource Inventory from the Ordered Systematic Procedure to a Probability Proportional to Size with Replacement (PPSWR) design. Schwarz (2000) documents the procedures to be employed for this revised design and this document summarizes the main report. The latter should be consulted for details and examples of using the proposed design. Neither this summary nor the main report revisits the question of the revised design against the old design - the reader should refer to Ott (1999), and Ott and Sit (1999).

2. Methodology

2.1 Sampling Frame and Sample Size

The first step in selecting polygons is to construct a **sampling frame** for the inventory unit. A sampling frame is simply a list of all the sampling units (the polygons) in the population of interest along with any other information available for each unit. In the VRI protocol, this list is available from the existing inventory files, or new photo-interpretation studies and contains such information as map #; polygon #, polygon area,

and photo-interpreted values such as BC land classification, species composition, stand height, age, density, etc.

Once the frame is established, two immediate questions arise: (1) how many polygons should be selected and (2) how should the sample be allocated among polygons with different attributes. We agree with the recommendation of Stauffer (1995, p.39) that "Current indications is that, initially, 100-400 ground-sampled clusters can be anticipated for each inventory unit." Stauffer (1995) also recommends that a sequential process be used where a smaller (say 30-50) samples are chosen, the data collected, and if precision is inadequate, further samples are selected.

2.2 Stratification

It is often advantageous to subdivide the population of interest into distinct groups (called strata), and then conduct separate surveys within each stratum. This pre-stratification (because it occurs before the sample is selected) is often done for a number of reasons:

- Administrative convenience.
- Specialized methods are needed for different strata.
- Domain specific estimates required.
- Better precision.

Decisions on stratification should be based on groupings that are useful to most end-users yet are easily tracked and administered. For example, non-vegetated polygons, which are valuable from the ecological point of view, but are less useful in terms of timber value, can be grouped into a separate stratum for special consideration. Pre-stratifying by leading species, or ecological classification may be most useful as both are related and influence the main variables of interest.

The sampling effort can be allocated among the strata in a number of ways. In proportional allocation, the total sampling effort is allocated among the strata proportional to the total number of polygons or total area of the polygons. In a fixed allocation, the total sampling effort is allocated among the strata to achieve some pre-

specified targets or to allocate effort in strata where the value of information is highest. More sophisticated allocations are also possible using if prior knowledge about the variability in each stratum is available

2.3 Selecting polygons

Once the strata are established and the number of polygons to be selected from each stratum determined, the selection process proceeds. Under this proposal, polygons are selected randomly, but the chance of selection varies from polygon to polygon.. Larger polygons are given a larger probability of selection. In particular, the probability of selection at each draw is proportional to the area of the polygon. Polygons are "put back into the sampling pool" after each selection so that every polygon has the potential to be selected on every draw.

It would be possible to modify the selection process by using other criteria other than area of the polygon to determine the sampling probabilities. The theory presented below is unchanged, but a poor choice of "size" variable can lead to estimates with extremely poor precision.

2.4 Selecting grid points within sample polygons

Ground points are selected within selected polygons using a fixed 100 m by 100 m grid, which is linked to the 20 km by 20 km provincial grid system. Once the polygons are selected, the 100 m by 100 m grid system is used to identify a set of grid points within each polygon. Because the initial polygons were chosen with replacement, the selection of points within a polygon can be done in any number of ways (as long as it leads to an unbiased estimate of the polygon total) and could differ among the polygons without introducing any additional complexity into the estimation process.

2.5 Estimators

This report discusses four basic techniques to obtain estimates of interest (say the total volume) for an inventory unit:

- inflation estimator - the Phase II information is used without reference to the Phase I information
- ratio estimator - a linear relationship through the origin between the Phase I and Phase II measurements is used to adjust the Phase I total.
- regression estimator - a linear relationship not restricted to pass through the origin is used to adjust the Phase I total
- geometric mean estimator - a linear relationship not restricted to pass through the origin and accounting for "errors" in both phases is used to adjust the Phase I total.

Estimates can be obtained for the total of an inventory unit, the mean per hectare of an inventory unit, or a proportion of the inventory unit bearing a certain attribute. Only estimating the total will be discussed here as the mean per hectare can be derived from the inventory unit total and the area of the inventory unit. Similarly, Cochran (1977) illustrates how to estimate proportions – the estimators for proportions are analogous to those used for estimating means.

Estimates for totals for part of an inventory unit - termed a domain - may be requested. For example, a total may be desired for the volume in a particular biogeoclimatic zone. These domains may cut across strata. The full report gives examples of how to determine estimates for domains.

Regardless of the method used to estimate the overall population total, the first step is to obtain estimate of the total value of the variable for each of the selected polygons. This is obtained from the data collected in the ground survey. Regardless of the number of ground points measured in a specific polygon, the only information needed is the estimate of the total value for that polygon, denoted $\hat{Y}_i$. For example if grid points

were selected using a simple random sample within a polygon, the estimated polygon total is found by multiplying the mean over the ground points by the polygon area.

Finally, it should be emphasized that the formulae for both the estimator and especially the estimated standard error of the estimator are tightly tied to the sampling procedure used to select the sample. If a different design from PPSWR is used, then the results of this paper are not applicable.

2.5.1 The Inflation Estimator

The simplest estimate of the population total uses only the estimated polygon totals from the selected polygons – the Phase I information is not used. This estimator is often called an **expansion** or **inflation** estimator because the estimated polygon totals are expanded or inflated to estimate the population total.

This is first done independently for each stratum in the population. For stratum h from which n_h polygons were sampled, the estimated population total for the stratum is computed using the Hansen-Hurwitz estimator as:

$$\hat{Y}_{h,\text{inf}} = \frac{1}{n_h}\sum_{i=1}^{n_h}\frac{\hat{Y}_i}{p_i}$$

and its estimated variance (the se^2) is found as:

$$v\left(\hat{Y}_{h,\text{inf}}\right) = \frac{1}{n_h}\frac{\sum_{i=1}^{n_h}\left(\frac{\hat{Y}_i}{p_i} - \hat{Y}_{h,\text{inf}}\right)^2}{n_h - 1}$$

The grand total (over all strata) and the estimated variance of the grand total are found as:

$$\hat{Y}_{\bullet,\text{inf}} = \sum_{h=1}^{H}\hat{Y}_{h,\text{inf}}$$

$$v\left(\hat{Y}_{\cdot,\mathrm{inf}}\right)=\sum_{h=1}^{H} v\left(\hat{Y}_{h,\mathrm{inf}}\right)$$

Note that the estimated variance of the grand total is found by adding the individual estimated variances of the stratum total and not by simply adding the individual estimated standard errors (i.e. the se^2 terms must be added rather than simply summing the *se* terms.).

Once the total is obtained either at the stratum level or the inventory unit level, it can be converted to a "per hectare" basis by simply dividing it by the area of the stratum or inventory units. The standard error and confidence limits for the "per hectare" measurement are found by also dividing the standard error or confidence limits for the total, by the total area. When applied at the stratum level, this leads to:

$$\hat{\bar{Y}}_{h,\mathrm{inf}}=\frac{1}{A_h}\hat{Y}_{h,\mathrm{inf}}=\frac{1}{A_h}\frac{1}{n_h}\sum_{i=1}^{n_h}\frac{\hat{Y}_i}{p_i}$$

$$v\left(\hat{\bar{Y}}_{h,\mathrm{inf}}\right)=\frac{1}{A_h^2}v\left(\hat{Y}_{h,\mathrm{inf}}\right)=\frac{1}{A_h^2}\frac{1}{n_h}\frac{\sum_{i=1}^{n_h}\left(\frac{\hat{Y}_i}{p_i}-\hat{Y}_{h,\mathrm{inf}}\right)^2}{n_h-1}$$

where A_h is the area of stratum h. Combining over all strata, this gives:

$$\hat{\bar{Y}}_{\cdot,\mathrm{inf}}=\frac{1}{A}\hat{Y}_{\cdot,\mathrm{inf}}=\frac{1}{A}\sum_{h=1}^{H}\hat{Y}_{h,\mathrm{inf}}=\frac{1}{A}\sum_{h=1}^{H}A_h\hat{\bar{Y}}_{h,\mathrm{inf}}=\sum_{h=1}^{H}W_h\hat{\bar{Y}}_{h,\mathrm{inf}}$$

$$v\left(\hat{\bar{Y}}_{\cdot,\mathrm{inf}}\right)=\frac{1}{A^2}v\left(\hat{Y}_{\cdot,\mathrm{inf}}\right)=\frac{1}{A^2}\sum_{h=1}^{H}v\left(\hat{Y}_{h,\mathrm{inf}}\right)=\frac{1}{A^2}\sum_{h=1}^{H}A_h^2v\left(\hat{\bar{Y}}_{h,\mathrm{inf}}\right)=\sum_{h=1}^{H}W_h^2v\left(\hat{\bar{Y}}_{h,\mathrm{inf}}\right)$$

where A is the total area of the inventory unit, and W_h is the relative size of stratum h. This gives the standard result that the mean for the entire inventory unit is a weighted average of the stratum means.

A similar conversion from an estimate of a total to an estimate of a mean can be done for any estimator or sampling design.

2.5.2 Ratio Estimators

In the ratio estimator, the estimated ratio between the Phase I and II total within the sample is used to adjust the Phase I population total. This is known as model-assisted sampling – the presumed relationship between the Phase I and Phase II totals is used to improve the estimation process (Särndal et al., 1992).

There are two forms using ratio estimators – these are commonly called separate and combined ratio estimators. The former is likely to be of most interest to forestry studies and consequently presented in this summary report.

In the separate ratio estimator, separate ratio estimates of the population total are formed for each stratum, and then the estimated stratum total are added together. One common model assumes that in each stratum the variance in the response is proportional to the Phase I total. Under this model, the estimated total for each stratum is found as:

$$\hat{Y}_{h,ratio} = \hat{R}_h X_h = \frac{\hat{Y}_{h,inf}}{\hat{X}_{h,inf}} X_h = \frac{\frac{1}{n_h}\sum_{i=1}^{n_h}\frac{\hat{Y}_i}{p_i}}{\frac{1}{n_h}\sum_{i=1}^{n_h}\frac{X_i}{p_i}} X_h = \hat{Y}_{h,inf} + \hat{R}_h\left(X_h - \hat{X}_{h,inf}\right)$$

There are several estimates of the variance of the estimator (Särndal et al. 1992; Section 7.2). One simple estimate is:

$$v\left(\hat{Y}_{h,ratio}\right) = \frac{1}{n_h}\frac{\sum_{i=1}^{n_h}\left(\frac{e_i}{p_i} - \frac{1}{n_h}\sum_{i=1}^{n_h}\frac{e_i}{p_i}\right)^2}{n_h - 1} = \frac{1}{n_h}\frac{\sum_{i=1}^{n_h}\left(\frac{e_i}{p_i}\right)^2}{n_h - 1} \quad \text{where } e_i = \hat{Y}_i - \hat{R}_h X_i$$

The grand total and estimated variance of the grand total are found as before:

$$\hat{Y}_{\cdot,ratio} = \sum_{h=1}^{H}\hat{Y}_{h,ratio}$$

$$v\left(\hat{Y}_{\cdot,ratio}\right) = \sum_{h=1}^{H} v\left(\hat{Y}_{h,ratio}\right)$$

The ratio estimators assume that the relationship between the response and auxiliary variable is linear through the origin. It performs well, if the variability in the response variable also increases with the auxiliary variable. The guidelines of Stauffer (1995, p.29-32) on model assessment and fit are also important.

2.5.3 Regression Estimators

In regression estimators, the relationship between the Phase I and Phase II polygon total is again assumed to be linear, but not necessarily passing through the origin. The basic concept is to estimate the regression line between the two phases and use the regression line to adjust the inflation estimator.

As in the ratio estimator, there are two possible forms of the regression estimator and only the separate regression estimator is presented here. The form of the regression estimators will depend upon the assumed population model (Sarndal et al, 1992; Chapter 7). Under a model where there is a linear relationship between the Phase I and Phase II values with a constant variance, the estimated intercept and slope for each stratum are derived following Särndal et al (1992, p.230, Remark 6.4.4) as:

$$a_h = \frac{\hat{Y}_{h,inf} - b_h \hat{X}_{h,inf}}{\frac{1}{n_h}\sum_{i=1}^{n_h}\frac{1}{p_i}}$$

$$b_h = \frac{\frac{1}{n_h}\sum_{i=1}^{n_h}\frac{X_i\hat{Y}_i}{p_i} - \frac{\hat{X}_{h,inf}\hat{Y}_{h,inf}}{\frac{1}{n_h}\sum_{i=1}^{n_h}\frac{1}{p_i}}}{\frac{1}{n_h}\sum_{i=1}^{n_h}\frac{X_i^2}{p_i} - \frac{\hat{X}_{h,inf}^2}{\frac{1}{n_h}\sum_{i=1}^{n_h}\frac{1}{p_i}}} = \frac{\frac{1}{n_h}\sum_{i}^{n_h}\frac{(X_i - \tilde{X}_h)(Y_i - \tilde{Y}_h)}{p_i}}{\frac{1}{n_h}\sum_{i}^{n_h}\frac{(X_i - \tilde{X}_h)^2}{p_i}}$$

where $\tilde{X}_h = \hat{X}_{h,inf} \Big/ \frac{1}{n_h}\sum_{i=1}^{n_h}\frac{1}{p_i}$, $\tilde{Y}_h = \hat{Y}_{h,inf} \Big/ \frac{1}{n_h}\sum_{i=1}^{n_h}\frac{1}{p_i}$.

The estimated slope and intercept are used to predict the total for each polygon and the estimated population total for stratum h is found by summing all predicted values:

$$\hat{Y}_{h,reg} = \sum_{i=1}^{N_h} (a_h + b_h X_i) = N_h a_h + b_h X_h \approx \hat{Y}_{h,inf} + b_h (X_h - \hat{X}_{h,inf})$$

There are several variance estimators (Särndal et al.1992) and one simple estimator has the form:

$$v\left(\hat{Y}_{h,reg}\right) = \frac{1}{n_h} \frac{\sum_{i=1}^{n_h} \left(\frac{e_i}{p_i} - \frac{1}{n_h} \sum_{i=1}^{n_h} \frac{e_i}{p_i} \right)^2}{n_h - 1} = \frac{1}{n_h} \frac{\sum_{i=1}^{n_h} \left(\frac{e_i}{p_i} \right)^2}{n_h - 1} \quad \text{where } e_i = \hat{Y}_i - (a_h + b_h X_i)$$

The grand total and estimated variance of the grand total are found as before:

$$\hat{Y}_{*,reg} = \sum_{h=1}^{H} \hat{Y}_{h,reg}$$

$$v\left(\hat{Y}_{*,reg}\right) = \sum_{h=1}^{H} v\left(\hat{Y}_{h,reg}\right)$$

2.5.4 Geometric Mean Regression Estimators

It is usually assumed in linear regression that the X variable is measured without error. In the VRI project, this is clearly not the case. Both the Phase I and Phase II measurements are only estimates of the true polygon value. In cases where both X and Y are subject to variation around the true value, estimates of the slope for the relationship between X and Y are known to be biased downwards (Berkson 1950), and cases are often made for "error in variable methods", of which the geometric mean regression is one example. The magnitude of the bias is related to the magnitude of the measurement error relative to the variation along the X axis. For example, if the standard deviation of the measurement error is 10% of the standard deviation of the values along the X axis, the slope is biased by a factor of about 1%. (Angleton and Bonham, 1995). Consequently, unless the error in the X values is an appreciable fraction of the variability in the X values, bias is negligible.

The geometric mean regression is often uncritically recommended for error-in-variable problems. However, many users are unaware of the potential inconsistencies in

the estimator. First, if no assumption is made about the relative size of the two measurement errors, no estimator of slope will perform well because the problem has more parameters than can be estimated from the data. Second, the geometric mean regression is also biased unless certain assumptions about the ratio of the two error variances to certain statistics is satisfied. This is rarely the case in practice. Third, if the ratio of the error variances is known (e.g. it seems plausible that the error variances could be equal for both variables), then better methods are available. Draper and Smith (1997), and Riggs et al (1978) offer a complete discussion.

Given these caveats, the estimated geometric slope and intercept for stratum *h* are found as:

$$b_{h,gmr} = \sqrt{\frac{b_{h,Y\ on\ X}}{b_{h,X\ on\ Y}}} \times \textit{sign of slope} \text{ and}$$

$$a_{h,gmr} = \hat{\bar{Y}}_{h,inf} - b_{h,gmr}\hat{\bar{X}}_{h,inf}$$

where $\hat{\bar{Y}}_{h,inf}$ *and* $\hat{\bar{X}}_{h,inf}$ are the mean per polygons based on the inflation estimators.

Then, as in the regression estimator, the estimated slope and intercept are used to predict the total for each polygon and the estimated population total for stratum *h* is found by summing all predicted values:

$$\hat{Y}_{h,gmr} = \sum_{i=1}^{N_h}\left(a_{h,gmr} + b_{h,gmr}X_i\right) = N_h a_{h,gmr} + b_{h,gmr}X_h = \hat{Y}_{h,inf} + b_{h,gmr}\left(X_h - \hat{X}_{h,inf}\right)$$

with a simple variance estimator of:

$$v\left(\hat{Y}_{h,gmr}\right) = \frac{1}{n_h}\frac{\sum_{i=1}^{n_h}\left(\frac{e_i}{p_i} - \frac{1}{n_h}\sum_{i=1}^{n_h}\frac{e_i}{p_i}\right)^2}{n_h - 1} = \frac{1}{n_h}\frac{\sum_{i=1}^{n_h}\left(\frac{e_i}{p_i}\right)^2}{n_h - 1} \text{ where } e_i = \hat{Y}_i - \left(a_{h,gmr} + b_{h,gmr}X_i\right)$$

The grand total and estimated variance of the grand total are found as before:

$$\hat{Y}_{\bullet,gmr} = \sum_{h=1}^{H}\hat{Y}_{h,gmr}$$

$$v\left(\hat{Y}_{\bullet,gmr}\right) = \sum_{h=1}^{H}v\left(\hat{Y}_{h,gmr}\right)$$

2.6 Confidence intervals

Regardless of the method used to compute the estimate and its estimated standard error, an approximate confidence interval can be found by recourse to the central limit theorem which states that regardless of the distribution of polygon values, estimates that are functions of sample means will typically follow a normal distribution in large samples. Hence an approximate confidence interval is found as:

$$estimate \pm z_{\alpha} se$$

where z_{α} is the appropriate percentage point of a normal distribution. For example, for 95% confidence intervals, $z_{\alpha} = 1.96$.

Caution should be exercised when sample sizes are "small" or responses have a large skewness, e.g.. a long right or long left tail. In these circumstances, the above estimated confidence interval may not reflect the actual probability of including the true value of the unknown parameter.

An alternative method to estimate the standard errors would be through the use of a bootstrap. In this procedure, the observed data are resampled with replacement to create a series of "new" samples; the estimates are obtained from this "new" sample; and the actual distribution of estimates over the set of "new" samples are used to determine the lower and upper confidence limits. Rao (1997, Section 3.1) and Särndal et al (1992, Section 11.6) describes the construction of the bootstrap for complex survey design.

3. Computer Software

Unlike the analysis of experimental design data, software for the analysis of survey data is not readily available, but several standalone packages are available. They range in price from free to several thousand dollars per year. A review of existing software and the reasons why specialized software is necessary for the analysis of survey data is provided by the Survey Methods Section of the American Statistical Association at:

http://fas-www.harvard.edu/~stats/survey-soft/survey-soft.html

Fortunately, SAS Version 8 now contains procedures for the analysis of survey data that are integrated into the SAS system. There are three procedures:

Proc SURVEYSELECT – to select samples
Proc SURVEYMEANS – for estimates using inflation methods
Proc SURVEYREG – for estimation using ratio and regression methods.

These new procedures in SAS use the same methodology as in other packages and have similar capabilities. The advantage of using SAS over a standalone package is that all the data management and analysis tools already present in SAS can be used in conjunction with these new procedures without having to learn a new package.

It is recommended that proper computer packages specially designed for the analysis of survey data (e.g. SAS V.8) be used to process the data rather than trying to implement the formulae in this report using, for example, a spreadsheet. The latter often leads to errors in transcribing the formulae and numerical problems for large datasets. A sample set of SAS programs was created to analyze the simulated example in the Appendix. These can serve as templates for the analysis of a real survey. Note that the estimated variances from SAS may differ slightly from those computed using the simple variance formula as SAS uses a more complex estimator.

4. Discussion

This report details an explicit multi-stage design whereby polygon are selected in the first stage with probability proportional to size with replacement (PPSWR), and ground points selected within polygons by any method that leads to unbiased estimates of polygon totals.

The advantages of this approach are:

(a) It gives greater weight to larger polygons as they have a greater impact upon the overall population parameters.

(b) It is flexible. Polygons can be added or removed from the survey design easily. The number of ground points sampled within a polygon can vary from design specifications without introducing additional problems.

(c) Methods used to sample ground points can vary among polygons. Allocation of number of sampling points among the sampled polygons can be varied to optimize overall precision requirements.

(d) Missing sampling points within polygons are easily accommodated as long at they occur completely at random.

(e) It allows estimates of precision to be easily computed regardless of complexities that may occur at the first or second sampling stage. These estimates of precision incorporate implicitly most sources of variation in the survey that occur.

(f) Computer software is available to assist in the analysis data collected from this design.

(g) Stratification of polygons by various attributes is explicit rather than implicit in previous methods.

(h) Allocation of sampling points among the strata is flexible to meet precision and other requirements rather than being fixed to ensure that every hectare receives and equal chance of selection.

It should be noted that the proposed methodology leads to algebraically identical estimators as proposed in previous reports when indeed every hectare has an equal probability of selection.

The potential disadvantages to this approach are, for the most part, minor:

(a) It is possible that polygons could be selected multiple times. However, this is not expected to occur very often given the large number of polygons in a typical inventory unit and the relatively small sample sizes selected.

(b) Computing polygon totals for variables such as age may seem odd - however, this was implicitly done in previous methods even though the simple formulae did not explicitly include polygon area or number of trees in the estimation formulae.

(c) The analyses are tied to the polygons. These polygons may change over time and they may not be compatible with the interests of all users. However, adjustments can be made to account for changes in polygon definitions or to obtain estimates for groups of polygons (domains).

5. References

Angleton, G. M. and Bonham, C. D. (1995). Least squares regression vs. geometric mean regression for ecotoxicology studies. Applied Mathematics and Computation, 72, 21-32.

Berkson, J.B. (1950). Are there two regressions? American Statistical Association Journal 45: 164-180

Cochran, W. G. (1977), Sampling Techniques, Third Edition, New York: John Wiley & Sons, Inc.

Draper, N.R. and Smith, H. (1998). Applied regression analysis, 3rd edition. New York: John Wiley & Sons, Inc.

Ott, P, and Sit. V. (1999). A Statistical Review of the B.C. Forest Inventory Audit Sampling Scheme and Estimation Methods. BC Ministry of Forests, Resources Inventory Branch.

Ott, P. (1999). A Comparison of two Sampling Designs for use in the B.C. Forest Inventory: Stratified Probability Proportional to Size with Replacement (PPSWR) and Multiple Pass Ordered Systematic (MPOS) Sampling. BC Ministry of Forests, Resources Inventory Branch.

Rao, J.N.K. (1997). Developments in sample survey theory: an appraisal. Canadian Journal of Statistics, 25, 1-21.

Riggs, D.S., Guarnieri, J.A. and Addelman, S. (1978). Fitting straight lines when both variables are subject to error. Life Sciences, 22, 1305-1360.

Särndal, C.-E., Swensson, B., Wretman, J. (1992). Model assisted survey sampling. Springer-Verlag: New York.

Schwarz, C. J. (2000). The Statistical Estimation and Adjustment Process using a PPSWR sampling design in the Vegetation Resource Inventory. Prepared for the BC Ministry of Forests, Resources Inventory Branch.

Stauffer, H.B. (1995). The Statistical Estimation and Adjustment Process for the British Columbia Vegetation Resource Inventory. Prepared for the BC Ministry of Forests, Resources Inventory Branch.

Appendix A: Example

A simulated population was created based upon data from the Sunshine Coast provided by Dr. M Penner on behalf of the Ministry of Forests.

It consists of approximately 46,000 polygons covering about 800,000 ha. The non-vegetated polygons (those with non-blank non-productive code or a blank leading species code) are of little interest in the VRI and hence will be deleted from the population of interest. The resulting population consists of about 36,000 polygons covering about 550,000 ha. Of the vegetated polygons, about 80% of the polygons are in the cedar-western-hemlock biogeoclimatic zone - a domain of specialized interest.

From a detailed investigation of the polygons, it was decided that three strata are appropriate based on leading species. These were *F,H,* and *Others*. A preliminary investigation showed that about 200 polygons should be selected. The allocation of samples will be about 80:70:50 for the three strata, based on an approximate proportional allocation by total polygon area.

Table 1 presents a summary of the estimates computed for both the entire inventory unit and the cedar-western-hemlock biogeoclimatic zone using one of the domain estimation methods. The volume/ha estimates are derived by dividing the estimated total volume (and its estimated se) by the area of the stratum or unit respectively.

Table 1 Summary of the estimates for the total volume over the inventory unit computed using the various methods on the simulated data.

	Total volume (millions of m^3) for the entire inventory unit		Volume per hectare (m^3/ha) for the entire inventory unit.[1]		Total volume (millions of m^3) for the CWH Domain[2]	
Stratum	Estimated Total	Est SE	Estimated volume/ha	Est SE	Estimated Total	Est SE
			Inflation Estimator			
F	90.95	6.11	427.5	28.7	89.0	13.9
H	78.58	5.29	414.8	27.9	50.7	7.6
Other	53.50	5.05	329.9	31.1	40.3	9.6
Total	223.03	9.53	395.2	16.9	180.1	18.5
			Ratio Estimator			
F	88.46	3.81	415.9	17.9	79.6	3.9
H	75.48	2.47	398.4	13.0	56.2	2.6
Other	48.36	4.52	298.2	27.9	34.7	4.4
Total	212.30	6.41	376.2	11.4	170.4	6.4
			Regression Estimator			
F	91.84	3.08	431.7	14.4	83.3	3.6
H	78.78	2.60	415.9	13.7	59.2	2.5
Other	61.38	3.34	378.5	20.6	45.6	3.4
Total	231.99	5.26	411.1	9.3	187.3	5.2
			Geometric Mean Regression Estimator			
F	91.65	3.06	430.8	14.4	82.8	3.0
H	79.32	2.74	418.7	14.5	58.7	2.7
Other	61.44	2.73	378.9	16.8	44.3	2.6
Total	232.44	4.93	411.9	8.7	185.5	4.9

[1] Total area of 564,306.4 ha with 212,713 ha in stratum F, 189,447 ha in stratum H, and 162,150 ha in stratum Other.
[2]The CWH Domain are the polygons identified as falling in the cedar-western-hemlock biogeoclimatic zone.

Because of the strong linear relationship between the Phase I and Phase II readings, it is not surprising that the precision of the ratio and regression estimators is improved compared to the simple inflation estimator. The regression estimator performs comparable to the ratio estimator as the fitted lines do tend to intersect the origin. There is

little advantage to going to a geometric mean regression - the se are comparable to those from ordinary regression. As noted earlier, unless the error in X is considerable, there is little bias in ordinary regression. This result for the geometric mean regression may not be generalizable to a real survey where the error in X may be larger.

The Statistical Estimation and Adjustment Process using a PPSWR sampling design in the Vegetation Resource Inventory.[1]

Carl James Schwarz
Department of Statistics and Actuarial Science
Simon Fraser University
Burnaby, BC V5A 1S6

10 February 2001

---

[1] This report was prepared for the Ministry of Forests, Government of British Columbia, Canada as part of contract 52MFR00015VS.

Summary

This report describes a proposed sampling design for the Vegetation Resources Inventory. It is based on a two-stage approach. In the first stage, polygons are selected with probability proportional to polygon area and with replacement. In the second stage, points are selected within polygons from the provincial grid using any number of sampling methods. Ground crews take measurements at these sampling points and these are then used to estimate inventory values either by themselves or in conjunction with polygon values available from aerial photographic interpretation. Various estimators were proposed including a simple inflation estimator, a ratio estimator, a regression estimator, and a geometric mean regression estimator. Finally, a detailed worked example based on the Sunshine Coast is used to illustrate the various methods. A set of Frequently Asked Questions (FAQ) about the implementation of this design is also provided.

The current survey methods select ground points using the paradigm of "every hectare in the province has an equal probability of selection". This leads to a very simple algebraic form for the estimators, but is inflexible in dealing with problems during survey operations and does not lead to estimates of precision with known properties.

In this report, this requirement is relaxed in favor of the explicit multi-stage design. The advantages of this approach are:

(a) It gives greater weight to larger polygons as they have a greater impact upon the overall population parameters.

(b) It is flexible. Polygons can be added or removed from the survey design easily. The number of ground points sampled within polygons can vary from design specifications without introducing additional problems. Methods used to sample ground points can vary among polygons. Allocation of number of sampling points among the sampled polygons can be varied to optimize overall precision requirements. Missing sampling points within polygons are easily accommodated as long at they occur completely at random.

(c) It allows estimates of precision to be easily computed regardless of complexities that may occur at the first or second sampling stage. These estimates of precision incorporate implicitly most sources of variation in the survey that occur.

(d) Computer software is available to assist in the analysis data collected from this design.

(e) Stratification of polygons by various attributes is explicit rather than implicit as in previous methods. Allocation of sampling points among the strata is flexible to meet precision and other requirements rather than being fixed to ensure that every hectare receives and equal chance of selection.

Table of Contents

1. Introduction 7
1.1 General description of the VRI 7
1.2 The proposed protocol for Phase II 8
1.3 Sources of variation 9
1.4 Sampling and non-sampling errors 11
1.5 Outline of this report 14
2. Selecting Polygons and related issues 15
2.1 The Sampling Frame 15
2.2 Sample size and Stratification 15
2.3 Allocation of sampling effort among strata 18
2.4 Issues in selecting a sample 21
2.4.1 Selecting polygons 21
2.4.2 What if a polygon is selected twice? 22
2.4.3 Can the sample size be increased after the sample is selected? 23
2.4.4 What if the polygons change definition between the time the sample was selected and the survey is completed 23
2.4.6 What if ground points are missing, changed, etc within polygons. 25
2.4.7 What is the sampling unit – a polygon or a point? 26
2.4.8 What if a domain has a very small sample. Can more polygons be added later? 26
2.4.9 What if not all selected polygons are sampled 28
2.4.10 How many points should be selected within each polygon? 28
3. Estimation 29
3.1 Estimating the total for each sampled polygon. 29
3.2 Estimates that use only the ground information 31
3.2.1 The Inflation Estimator 31
3.3 Estimates that use the relationship between the Phase I and Phase II values. 37
3.3.1 Ratio Estimators 38

3.3.2 Regression Estimators 42
3.3.3 Geometric Mean Regression Estimators 44
3.4 Confidence intervals 47
3.5 Adjusting individual polygon values 48
4. Computer Software 49
5. Example 50
5.1 The Phase I data 50
5.2 Deciding upon strata, sample sizes, and allocation 53
5.3 Selecting the sample 56
5.4 Obtaining the Phase II data 56
5.5 The Inflation estimates 57
5.6 The Ratio Estimates 59
5.7 The Regression Estimates 59
5.8 Geometric Mean Regression Estimators 60
5.9 Summary of the estimates 60
6. Discussion 64
7. Further research 66
8. References 67
Appendix C Frequently Asked Questions (FAQ) 74
C.1- Why can't I use the simple SRS formula for the estimates and standard errors? . 74
C.2 What is the effect of fixed grid system? 76
C.3 What is the sampling unit – a polygon or a point? 77
C.4 What happens if I can't sample all points within a polygon? 78
C.5 Why do I use only the estimated polygon totals and not the "point" value. Aren't I throwing away information? 78
C.6 Is it a problem that the aerial values are subject to "error" 81
C.7 Isn't it better to stratify the population as much as possible? 81
C.8 How many points should be sampled in each polygon? 83
C.9 What if the polygon boundaries change between the time the sample is selected and measured? 83

C.10 Is it sensible to stratify by total volume as this is the primary attribute of interest? 84
C.11 Why don't the estimators include stratum weights? 85
C.12 Why don't the variances formulae account for variability in polygon areas? 86
C.13 How is choice made between the inflation, ratio, or regression estimator? 87
C.14 What is the difference between a domain and a stratum or post-stratum 87
C.15 When should Method 1 and Method 2 be used for domain estimation? 88

1. Introduction

1.1 General description of the VRI

The Vegetation Resource Inventory (VRI) is a provincial survey that collects data on the type, amount, and location of vegetation in British Columbia. The survey is carried out in inventory units, which may be as small as a watershed or as large as a Forest District. The survey process consists of two steps. In Phase I, aerial photographs are taken of the entire inventory unit. These are used to divide the inventory unit into homogenous areas called polygons and to provide photo-interpreted estimates for every polygon in the inventory unit. The second step of the survey process selects a sample of polygons from which ground location are selected for taking ground measurements. These ground measurements are then used to adjust the estimated inventory totals from Phase I accounting for the observed data from the ground points.

The second step in the proposed VRI sampling design is an example of a two-stage design. In the first stage, polygons will be selected (possibly after stratification). In the second stage, ground points from a fixed provincial grid that lie within the polygon will be selected and field crews will obtain actual measurements at ground level.

This general two-stage design has a number of advantages. First, the relatively fewer polygons at the first stage reduce data handling complexities involved in stratifying points, selecting points, and locating the sampled points. Second, there is a great deal of flexibility in how the points at the second stage can be selected and measured. The number of points selected from each polygon can be controlled to achieve cost and precision goals independently of which polygons are selected. For example, the number of points selected within each polygon can vary depending upon the type of polygon, the heterogeneity of the forest within the polygon, etc. As well, local problems, such as being unable to sample a particular point within a polygon because of safety reasons, can be

handled without compromising the design at the polygon level (refer to Penner (2000) for more details).

1.2 The proposed protocol for Phase II

The proposed VRI protocol for Phase II calls for selecting the polygons using a method called Probability Proportion to Size with Replacement (PPSWR).

This method has number of components. First, selection of the polygons to be sampled is done randomly, but the chances of selecting a polygon are not equal. Rather, larger polygons are given a larger probability of selection, in particular, the probability of selection at each draw is proportional to the area of the polygon. For example, a polygon which has an area of 1000 ha will have twice the probability of selection at each draw compared to a polygon whose area is 500 ha. The rationale for the selection probabilities being proportional to the size (area of the polygon) is that larger polygons contribute more to the overall total in the inventory unit and should be given more weight in the sample selection process and estimation of the total. Cochran (1977, Section 9A.5) indicates that when comparing equal probability selection methods and pps methods, pps methods are to be preferred when the mean per unit is unrelated to the size of the unit and that equal probability selection methods are to be favored when the unit total is unrelated to the size of the unit. In the VRI, the mean per unit could refer to the volume/ha. This is unlikely to be related to the polygon area. The unit total would be the volume for the entire polygon (volume/ha times polygon area) and this is likely related to the size of the unit. Consequently, pps sampling has strong theoretical advantages.

Second, polygons are "put back into the sampling pool" after each selection so that every polygon can be selected on every draw. The rationale for the "with replacement" aspect is that this gives a great deal of flexibility in how points can be chosen within each polygon and greatly simplifies estimation of the precision (the standard error) of the estimates. By choosing the initial polygons with replacement, the selection of points within a polygon can be done in any number of ways and could differ

among the polygons without introducing any additional complexity into the estimation process. All that is important is the method chosen to select points within each polygon gives an unbiased estimate of the polygon total.

Third, ground points are selected within the selected polygons. The province of BC has been overlaid with a fixed 20 km by 20 km grid system. Once the selected polygons are established, a 100 m x 100 m grid system is overlaid on the larger grid system to identify a set of grid points within each polygon. Because the initial polygons were chosen with replacement, the selection of points within a polygon can be done in any number of ways and could differ among the polygons without introducing any additional complexity into the estimation process. For example, points within a polygon could be chosen systematically, or using a simple random sample. For example, if a simple random sampling of grid points is to be used, the grid points within a selected polygon are numbered, say 1, 2,... K. Then to select a point from the set, a random number from 1,...,K are chosen and the corresponding grid point is scheduled to be surveyed. The overall probability of selection of any grid point in the province is then equal to the product to selecting the surrounding polygon and selecting that point within the polygon, i.e. $P(\textit{select a point}) = \frac{\textit{area of polygon}}{\textit{total area}} \times \frac{1}{\textit{number of grid points in polygon}}$

Previous proposed designs for the VRI tried to ensure that every point in the province had the same probability of selection. This constraint made the sample selection and estimation process unnecessarily complicated. By adopting a two stage design, the surveyor has a great deal of flexibility in allocating sampling effort and conducting the survey. It also readily adapts to problems in the field such as missing polygons or points not being able to be sampled.

1.3 Sources of variation

It is self evident that not every point in BC is identical. The variation among the points in BC can be broken into several "sources" of variation at various levels:

- Inventory unit variation.

- Strata within inventory units
- Polygons within strata
- Points within polygons
- Grid-to-grid variation if the provincial grid was moved
- Measurement errors that occur when taking measurements at selected points.
- Measurement errors in determining the sizes and outlines of polygons

Any sampling process must recognize and account for these sources of variation either implicitly or explicitly.

Inventory unit variability is accounted for implicitly by conducting a separate survey within each inventory unit. The results from one inventory unit are not extrapolated to another inventory unit.

Within inventory units, areas are delineated into polygons based upon photo-interpretation. These polygons are drawn to enclose an area of land that is as homogeneous as possible. For example, a polygon may be drawn around a large stand of trees all approximately the same age in the same type of soil and elevation.

During the survey process, these polygons may be grouped into larger collections, called strata, for either administrative convenience or to obtain greater precision in the estimates. For the latter objective, polygons within strata should be as homogeneous as possible while the strata themselves should be as different as possible. For example, strata could be defined by leading species.

It is impossible to take exhaustive measurements on the entire polygon to completely census its tree population. Rather, a random selection is made from a set of grid-points enclosed within the polygon. Variation among the grid-points within a polygon would measure the point-to-point on the grid variation. There is a separate VRI process called within polygon variation (WPV) that was supposed to capture and give an

indication of the magnitude of this variation. This could be used in planning the allocation of effort among and within polygons.

Even at these grid points, measurements that are taken of the forest attributes are not perfect – measurement error is present. Presumably, replicated measurements taken at a grid point could be used to estimate the measurement error. Information on the likely magnitude of the error may also be available from other studies.

Consequently, nearly all but grid-to-grid and polygon area error variation can be quantified in this survey. The grid-to-grid variation cannot be estimated because no measurements are taken off the grid! Fortunately, it is expected that any biases caused by the fixed grid and any contribution of the fixed grid to the overall variation is expected to be small and will be neglected. Polygon area error cannot be quantified because there is no objective way to determining any "true" size for a polygon. It is theoretically possible to have several interpreters draw polygon boundaries and then use these to get a feel for the likely magnitude of the errors. Fortunately, it is expected that such errors will not cause bias in the estimates and variation from this source of error will be captured in the estimated precision (see C.13 for more details).

1.4 Sampling and non-sampling errors

A complete census and measurement of every tree in the province could in theory give an exact value for any attribute. This is obviously impossible to obtain. Consequently, surveys must be conducted where only a small number of individual elements from the population are surveyed and the results from this survey are extrapolated to give an estimate for the entire population.

Because only a small fraction of the population is actually measured, the estimates so obtained will not equal the true population value exactly. Rather, if the survey could be conceptualized as being repeated over and over, some of the estimates will be larger than the true population value and some of the estimates will be smaller than the true

population values. The deviations of the estimates from the true population value are known as sampling error.

Statistical theory can quantify the magnitudes of the sampling errors if a **random sample** of elements is chosen. The randomization process in selecting units for the survey is what allows estimates of the likely size of the sampling error to be computed. Presumable, a larger sample gives "better" results, i.e. it seems intuitive that a larger sample should have, on average, a smaller sampling error than a small sample. However, this is true only if both samples were random samples. In these cases, it is said that the sample size control the precision of the estimates.

Survey methodology can only control uncertainty due to sampling errors. Other errors are possible in other survey that cannot be quantified through the sampling process. For example, if a measuring device always reads 10% too low, this would never be known from the sampling methodology and any statements about the precision of an estimate would ignore this non-sampling error. Consequently, every aspect of the survey must be carefully examined to avoid introducing any non-sampling errors into the results.

In some cases, some of the non-sampling errors can also be estimated. For example, in the VRI, a sub-sample of the selected ground points is selected and professional foresters are sent to the selected points to survey the exact same locations. The difference between the "exact" reading from the foresters and the earlier readings can be used to "adjust" the first ground measurements. This is an example of a "2-phase survey", where a sample of a sample is selected, but the analysis of these types of surveys is beyond the scope of this report.

Note that it is often assumed that the ground estimates are "better" estimates of the polygon attribute than the photo estimate. For some of the attributes (e.g., species composition), the photo interpreter looks at the entire polygon and summarizes it. This may not be a personalized attentive look at each tree in the polygon but certainly obvious

anomalies would be seen that be would overlooked when at an actual ground location. Consequently, in some cases, the photo estimate provides a better estimate for the polygon than the ground sample. This "non sampling" error is also very difficult to quantify.

The precision of an estimate is often expressed numerically using the "variance of the estimator", the "standard error of an estimator", the "coefficient of variation of an estimate", or a "confidence interval for the true population value". All of these concepts are related to each other and are interchangeable, i.e. if any one of the four is known, the other three can be derived but conversions to or from a coefficient of variation will also require knowledge of the mean.

Perhaps the most difficult concept in statistics to grasp is that of a sampling distribution. Conceptually, every sample taken from a population will give a different estimate. The sampling distribution of the estimator is simply the entire set of possible estimates if every possible different sample was taken. Some of the estimates will be larger than the true population value; some will be smaller than the true population value. If the average value of the estimate over all possible samples is equal to the true population value, the estimator is said to be **unbiased.**

The **variance of the estimator** is simply the variation of these possible estimates taken over all possible samples and the **standard error** (abbreviated "se") is the square root of the variance. [The term "standard error" is unfortunate as it denotes a "mistake", but for historical reasons, this is the preferred term]. The **coefficient of variation of an estimator** (denoted cv) is simply the ratio of the standard error to the average value of the estimate. All three terms are an attempt to quantify the amount of uncertainty caused by taking only a sample.

Lastly, the uncertainly about the true population value is often expressed by a **confidence interval** for the value. Typically, statements will be made such as "a 95%

confidence interval is from 10 to 20". The actual technical interpretation of this statement is not very illuminating, but roughly speaking the survey method will create an appropriate interval such as the above that will contain the true value 19 times out 20. [Unfortunately, unless Bayesian methods are being used, it is not technically correct to say that there is a 95% chance that the true population value is between 10 and 20.] Most 95% confidence intervals in survey methodology are approximately of the form

$$estimate \pm 2 \times se$$

where the factor "2" is related to the confidence level (95%) desired. Consequently, from the above confidence interval, it can be deduced that the "2 se" term is approximately 5, and it is technically correct to state that "the results will be accurate to within 5 units, 19 times of 20" as an interpretation of the results. – in other words, the true unknown population value is unknown, but there is a 95% chance that the procedure will give an estimate that within 5 units (either above or below) of the true value.

Lastly, the actual variance, standard error, coefficients of variation, or the confidence interval must also be estimated from the data at hand. Consequently, a distinction is often made between the *theoretical* variance and the *estimated* variance – although in practice the distinction is often glossed over.

1.5 Outline of this report

This report will examine many of the issues that will likely arise in implementing the proposed VRI PPSWR protocol. It will discuss issues related to the selection of the polygons, stratification, and allocation of sampling effort among the strata. The basic estimators used for this design will be outlined. Finally, a detailed worked example using simulated data based on the SunShine Coast will be used to illustrate the sample selection, allocation, and estimation procedures using the SAS programming system.

2. Selecting Polygons and related issues

2.1 The Sampling Frame

The first step in selecting polygons is to construct a **sampling frame** for the inventory unit. A sampling frame is simply a list of all the sampling units in the population of interest along with any other information available for each unit. In the VRI protocol, this list is available from the Phase I studies and contains such information as the size (area) of the polygon, and photo-interpreted values such as leading species, estimated timber volumes etc.

At this stage, the exact population on interest should be carefully defined. For example, are non-vegetated polygons or vegetated non-treed polygons of any interest? If not, these should be excluded from the frame and not included in the sample selection or analysis methods. Similarly, the completeness of the frame should also be considered – are the polygon boundaries and definitions still appropriate? Many of the problems of incomplete frames encountered in survey of human populations are not likely to be much of a problem except perhaps in remote areas where aerial photography may not be complete.

2.2 Sample size and Stratification

Once the frame is established, usual practice would be to select a sample from the units in the frame using some design and then to proceed to estimate the parameter of interest.

Two immediate questions arise: (1) how many polygons should be selected and (2) how should the sample be allocated among polygons with different attributes.

> The recommendations of Stauffer (1995, p.39) are likely appropriate, i.e.
> "Current indications is that, initially, 100-400 ground-sampled clusters can be anticipated for each inventory unit. With pre- and post-stratification, much smaller

sample sizes will apply per stratum. I would recommend, however, that minimum samples sizes n=30, and preferably, n=50, be used for the strata, where separate regression estimates are required, to provide adequate data to fit the regression model".

Stauffer (1995) also recommends that a sequential process be used where a smaller (say 30-50) samples are chosen, the data collected, and if precision is inadequate, further samples are selected.

The major problem is that sample size determination depends upon population information is not available, i.e the Phase II readings for every polygon. However, it seems reasonable that the Phase I information should provide some guidelines on the required sample sizes.

Following Cochran (1977, p.253), the theoretical variance of the inflation estimator (see section 3.2) ignoring any stratification is found as:

$$V\left(\hat{Y}_{inf}\right) = \frac{1}{n}\sum_{i=1}^{N} p_i \left(\frac{Y_i}{p_i} - Y\right)^2$$

where $p_i = a_i/A$ and Y is the total Phase 1 value. The summation term can be computed from the Phase I information, and various sample sizes (n) tried until the cv

$$CV\left(\hat{Y}_{inf}\right) = \frac{\sqrt{V\left(\hat{Y}_{inf}\right)}}{Y}$$

reaches acceptable levels.

However, in many cases, it is advantageous to subdivide the population of interest into distinct groups (called strata), and then conduct separate surveys within each stratum. This pre-stratification (because it occurs before the sample is selected) is often done for a number of reasons:

- Administrative convenience. For example, there may be different personnel available to sample, or the units may cover a wide geographic area and it is more convenient to break the problem into more local districts.

- Specialized methods are needed for different strata. For example, sampling ground points in different types of forest cover may require specialized equipment. Consequently, polygons needing different sampling methods may be grouped together.
- Domain specific estimates required. Separate estimates may be required for different parts of the population. For example, polygons may be stratified by leading species; or by old growth vs. 2nd growth areas. It is also possible to obtain domain estimates after the sample is selected without pre-stratifying, but the sample size in the domain is no longer under the control of the researcher.
- Better precision. In some cases, stratifying units into more homogeneous groups prior to selecting samples can results in a more precise estimate than sampling from the population as a whole. For example, stratifying by vegetated or non-vegetated would likely allow an increase in precision when estimating timber volume as the non-vegetated areas likely have no timber and taking a simple average over both types of polygons would introduce needless variability into the estimate.

"Stratification" can also be imposed upon the sample after-the-fact (denoted post-stratification). This could be useful if the aim is to increase the precision of the estimates. However, in many cases, the aim is to simply obtain estimates for domains within the population – this is more properly called "domain estimation" rather than stratification. The biggest disadvantage to post-stratification and domain estimation is that sample sizes are now random, i.e. it is impossible to predict in advance the sample size for a post-stratum or a domain. However, if the initial sample size is relatively large and the post-stratum or domain is not too rare, then the actual sample size is likely to be also relatively large and this may not be an issue. For example, if the initial sample asked for 100 polygons and the domain of interest (say a specific leading species) occurs in about 1/2 of the polygons, the observed sample size will be about 50.

However, stratification does come at a cost – increased complexity in the survey, and increased complexity in responding to post-survey ad-hoc requests. For example, suppose that a review of the study objectives leads to a decision to pre-stratify by environmentally sensitivity and by three different leading species classes giving a total of 6 initial strata. If the survey also wanted to stratify by old growth vs. 2nd growth, this would lead to a total of 12 strata with the administrative complexity of dealing with essentially 12 independent surveys. Then if a post-survey request wanted to examine density effects (say low, medium or high number of stems), a total of 36 possible post-strata would have to be created many of which may have very small or zero sample sizes.

Consequently, the decisions on stratification should choose strata that are useful for the most number of end-users as possible yet are easily tracked and administered. Some initial screening should also be done when the population is being defined. For example, non-vegetated polygons are likely not very interesting and should be removed from the frame. Pre-stratifying by leading species, or ecological classification may be most useful as both are related and likely influence the main variables of interest to many users.

2.3 Allocation of sampling effort among strata

The allocation of sampling effort among the strata can be done in a number of ways in pre-stratification.

In proportional allocation, the total sampling effort is allocated among the strata proportional to the total number of polygons or total area of the polygons. For example, suppose there were two strata where stratum 1 contains 1000 polygons with a total area of 2 million ha, and stratum 2 contains 2000 polygons with a total area of 1 million ha. Allocation proportion to the number of sampling units would allocate 1/3 of the effort to stratum 1 and 2/3 of the effort to stratum 2. Allocation to the area of each stratum would allocate 2/3 of the effort to stratum 1 and 1/3 to stratum 2.

One advantage of proportional allocation is that the final estimators have a very simple form for computation purposes (they are known as self weighting estimators). However, in this era of easily available computer software, there is no real advantage to self weighting estimators.

In a fixed allocation, the total sampling effort is allocated among the strata to achieve some pre-specified targets or to allocate effort in strata where the value of information is highest. For example, in most surveys where the sampling fraction (the proportion of units actually sampled from the population) is small, the precision of estimators is almost directly proportion only to the sample size and independent of the population size. Consequently, allocating roughly equal sample sizes to all strata regardless of the number of polygons in each stratum would achieve roughly equal precision in all strata. Or, certain strata may be known to be not of interest (e.g. non-vegetated) or of limited interest (environmentally sensitive areas where logging is not allowed). Consequently, the number of samples allocated to these strata should be small as the value of the information is also small.

More sophisticated allocations are also possible. For example, if prior knowledge about the variability in each stratum is available (e.g. from the Phase I data), then sampling efforts can be allocated among the strata to obtain the best possible overall precision possible. In these cases, strata that are larger or more variable tend to receive more sample effort to compensate for their larger effect upon the overall total and for the larger variability within the stratum. In a similar fashion, if the relative costs of sampling in the various strata also differ, this can be used to allocate samples to obtain the best precision for a given cost. Cochran (1977) reviews this in more detail and an example is presented in Section 5.2.

One of the primary advantages to pre-stratification is that sampling effort can be deliberately distributed among the strata to achieve a number of objectives. In post-stratification, this is not possible – the sample size in each stratum is random and cannot

be controlled. For this reason pre-stratification is to be preferred if the important strata can be identified in advance.

Similarly, the allocation of sampling effort among the pre-strata should be considered carefully. Again, it is most sensible to allocate most sampling effort to vegetated polygons and very little to non-vegetated polygons regardless of the number or total area of the polygons in each stratum. As the VRI is supposed to provide an ecologically based inventory, leading species may serve as a surrogate for local ecological factors (soil type, rain fall etc). Consequently, stratification by leading species may also be suitable.

The allocation of the sample among the strata can be examine in more detail again using the Phase I information as a surrogate to what will happen in the real population. Following Cochran (1977, p.253), the theoretical variance of the inflation estimator within each stratum h is:

$$V\left(\hat{Y}_{h,inf}\right) = \frac{1}{n_h}\sum_{i=1}^{N_h} p_i\left(\frac{Y_i}{p_i} - Y_h\right)^2$$

with $p_i = a_i/A_h$ and Y_h the stratum total from the Phase I information. The summation term within each stratum can be computed as:

$$Var_h = \sum_{i=1}^{N_h} p_i\left(\frac{Y_i}{p_i} - Y_h\right)^2$$

and the total variance for a particular allocation of n_h to each stratum is found as:

$$Var_{total} = \sum_{h=1}^{H} \frac{Var_h}{n_h}$$

Using the above, the overall precision for different allocation methods can be investigated. Numerical methods can also be used to investigate the optimal allocations to minimize the overall variance. This can be easily done using a spreadsheet program once the summation terms are collected and is illustrated in Section 5. In many cases, a ratio or regression estimator may be used. Unfortunately, without an initial sample, it is impossible in advance to investigate the effects of different sample allocations upon

precision. It seems most sensible to use the simple inflation estimator as a rough guide to sample size and allocation keeping in mind that more complex estimates may give better precision.

Some general guidelines can also be drawn up. First, stratification leads to gains in precision if the polygons within strata are homogeneous with respect to the attribute of interest (e.g. volume/ha) but the strata are as different as possible. However, it is not advantageous from an operational point of view to stratify too finely as each stratum will require its own sampling frame, its own randomization procedure, etc. Second, theoretical consideration (Cochran, 1977, p. 99) shows that if the polygons within strata are approximately equally variable in all strata, then a proportional allocation (in this case by area) will capture most of the possible theoretical gains in precision compared to a fully optimal allocation. These two points are illustrated in the worked example of Section 5.

2.4 Issues in selecting a sample

2.4.1 Selecting polygons

Polygons are to be selected using a PPSWR design where the size variable is the area of the polygon. Presumably, larger polygons have a greater influence upon the overall total than smaller polygons and should be given a greater chance of selection. For example, suppose that a stratum contained 1999 polygons each of 1 ha and 1 polygon of size 1000 ha. It would not make much sense to give the smaller polygons the same chance of selection as the larger polygon.

Once the areas are available for each polygon within each stratum, some preliminary tabulations should be constructed to review the range of areas available and to check for obvious errors in the database. For example, are there any polygons with areas of 0 ha? Are there any polygons with areas that seem excessive?

The probability of selection for each polygon in any specific draw (p_i) is then determined by the ratio of the polygon area to the total area in that stratum, i.e.

$$p_i = \frac{a_i}{A_h}$$

where a_i is the area of polygon i and A_h is the total area of all polygons in stratum h.

In actual practice, selection of polygons will likely be done using a computer rather than by hand. Consequently, details on the actual selection process are not presented here; but refer to the detailed example later in this report.

It would be possible to modify the selection process by using other criteria other than area of the polygon to determine the sampling probabilities. The theory presented below is unchanged, but a poor choice of "size" variable can lead to estimates with extremely poor precision.

2.4.2 What if a polygon is selected twice?

Because the selection of polygons occurs with replacement, it is possible that the same polygon could be selected twice. However, given the large number of polygons within a stratum (several thousand) and the relatively small number of polygons to be selected (in the low hundreds), the number of times that this will occur is very small.

In theory, if a polygon is selected twice, it should be measured independently twice and occur twice in the sample. It is possible to derive more complex estimators that only use the unique polygons (Pathak, 1962), but these are too complicated to be of any use in practice. If duplicated polygons are measured only once, or if a substitute polygon is selected, this does have some theoretical disadvantages as the design is no longer with replacement. However, given the relatively rare expected occurrence of this event, any "errors" introduced into the estimates is expected to be small relative to the overall precision of the estimate.

For the PPWSR design, duplicate polygons are treated as if they were distinct polygons, i.e. ground points are selected for each selection independently of the ground points for the duplicate selections; data is entered for this polygon twice (using the different set of ground points), etc. [Of course, operationally, all the ground points for a selected polygon would be measured at the same time.] Some care will need to be taken so that the ground points can be associated back to their respective 'selections' of the same polygon.

2.4.3 Can the sample size be increased after the sample is selected?

In some cases, additional resources may be available after the initial sample has been selected and additional polygons can be added to the sample. Because the design is a with replacement design, simply select additional polygons according to the protocol in Section 2.2.1. No complications are introduced into the estimation process other than adjusting the sampling weights (see Section 5.5) to reflect the new sample size.

2.4.4 What if the polygons change definition between the time the sample was selected and the survey is completed

Polygons are defined to be relatively homogeneous by forest type and other attributes. These were presumably based on aerial photographs. However, the polygon could cange between the time it was defined and actually sampled. For example, a fire could affect the trees within the polygon; the polygon could be harvested; or the original information that defined the polygon could be out of date.

The treatment of such changes is very complex. Here are several example illustrating a range of complexity that could occur:

- Suppose that the inventory unit was initially stratified by leading species and a polygon was initially classified in a certain stratum. When the ground surveys take place, it is found that in fact, the initial classification is wrong and that

the polygon belongs to another stratum. Both the original stratification of the frame and the sample have to adjusted to reflect these changes.

- Suppose that a polygon was originally coded by leading species but when the ground survey is performed, it is found that a fire has burned the polygon. The simple solution would be use the polygon as it but with a 0 value for the volume present. This will still lead to unbiased estimates of the total, but with worse precision. A post-stratification on the non-burned polygons would presumably lead to more precise estimates.
- Suppose polygon definitions are based on a photo-interpretation dated 20 years ago. While the ground points are being surveyed, a newer photo-interpretation takes place and all polygon boundaries change. Should the new boundaries be used in forming the estimates or should the estimates be based on the old boundaries? What should happen if two points selected within a single polygon based on the old boundaries now fall into two separate polygons based on the new boundaries?

A possible solution for some of these problem may be to reweight the sample data - however, it is not clear if this is suitable for all cases and further research is needed.

2.4.5 What if some polygons can't be sampled in their entirety?

Two cases of "missing values" should be distinguished. First, some polygons may not be accessible in their entirety for some reason. Second, a polygon may be accessible, but of the multiple points within the polygon, some of the individual points are not accessible. The latter is covered in the next section.

Missing observations almost always occur in surveys for many reasons. However, missing values can compromise the quality of the survey results if the polygons for which data were collected are different from the polygons where data are missing with regard to a outcome variable. For example, some polygons may be 'left for later' because they in some sense are 'difficult'. If polygons are measured sequentially in a certain pattern

(determined by logistic or otherwise to be convenient) then this secondary selection process has unknown ramification and most certainly cannot be considered to be completely at random. More advanced methods may be available for these cases that impute missing values with acceptable values - these are beyond the scope of this report,

If it can be assumed that polygons are missing completely at random, i.e. occurred just by chance and is not related to the variables being measured the missing polygons are simply omitted from the sample. Then, as Rao (1997, Section 4.1) indicates, unit non-response is usually handled by adjusting the weights of all respondents by a common non-response adjustment factor (see Section 5.5) to reflect the new sample size.

2.4.6 What if ground points are missing, changed, etc within polygons.

It may turn out, that while a particular polygon is accessible, not all ground points within the polygon are accessible and cannot be sample.

Because the sampling designs for selecting polygons was with replacement, the actual process of selecting ground points is quite flexible. Penner (2000) details procedures to follow when some ground points cannot be sampled. Note this concern is different then the previous section where entire polygons were inaccessible.

The essential point is that neither the number of ground points sampled within a polygon nor the method by which these are selected make any difference to the estimation process as long as the estimated value for the polygon total is statistically unbiased (see Section 3.1). Consequently, if was decided initially to sample three ground points within the polygon, but only two could be sampled, there would be no difficulties as long as the missing ground point was completely at random.

2.4.7 What is the sampling unit – a polygon or a point?

There are two sizes of sampling units in the Phase II protocol (or for that matter in any two stage design). The larger sized unit is the polygon. Polygons are selected at the first stage – the whole notion of ground points within the polygon are irrelevant at this point.

In the second stage, the sampling units are the ground locations within the selected polygons. In theory, there are an infinite number of ground points within a polygon (as points have no size), however, a frame is established within each polygon consisting of the points on a 100 m by 100 m grid that is aligned with the 20 km by 20 km provincial grid. This fixed grid system implies there are only a fixed number of possible sampling locations and the polygons can be regarded as having been divided in 1 ha squares centered on each grid point. Observations made from each sample grid node can then be regarded as belonging to the corresponding hectare. Obviously, this is not perfect, because the 1 ha squares will not, in general, fit properly into a polygon, but for most polygons, this would seem to provide a reasonable frame for sample selection. As noted earlier, one of the advantages of with replacement designs at the first stage is the complete flexibility in choosing points within the polygons – as long at the resulting values are unbiased for the polygon totals.

2.4.8 What if a domain has a very small sample. Can more polygons be added later?

The determination of sample size and allocation among the strata can lead to situation where the sample size in certain domains investigated after the study is complete is insufficient. For example, a certain domain (identified after the study was complete) may be relatively rare (e.g. 5% of the polygons) and so even if a large number of polygons were allocated, it is likely that only about 5% of the sample will fall into this domain.

There are two options to dealing with small domains.

First, if rare domains can be identified in advance of the survey, it is advantageous to create a separate stratum for this domain and allocate sufficient samples to the domain to ensure that results are sufficiently precise.

Second, if a domain is identified after an initial survey is completed, then a second survey can be constructed targeting only this domain. This second survey can use either a different design or the same design as the first survey. Then the estimates for the domain from the two surveys are easily combined as the two estimates are independent of each other. For example, let $\hat{Y}_{survey1}$ and $\hat{Y}_{survey2}$ be the estimates for the domain of interest from the two surveys with corresponding standard errors. A weighted average of the two estimates can be constructed

$$\hat{Y}_{weighted} = w_1 \hat{Y}_{survey1} + w_2 \hat{Y}_{survey2}$$

with an estimated standard error of:

$$se\left(\hat{Y}_{weighted}\right) = \sqrt{w_1^2 se^2\left(\hat{Y}_{survey1}\right) + w_2^2 se^2\left(\hat{Y}_{survey2}\right)}$$

The optimal weights (i.e. that give the smallest overall estimated se) are found as:

$$w_i = \frac{\frac{1}{se_i^2}}{\frac{1}{se_1^2} + \frac{1}{se_2^2}}$$

i.e. are related to the inverse of the variance of each estimate.

If the second survey had exactly the same design and targeted the same population as the first survey, then this is equivalent to simply adding additional samples to the initial survey and the data can be combined together into one larger sample. However, the two designs must be identical - any deviation to target a specific domain implies that the previous method to combine the estimates must be used.

It should be pointed out that collecting additional information on a single domain may introduce some complications in other parts of the analysis. For example, if this domain has a different relationship between the Phase I and II data than the other domains

in the population, which relationship should be used to adjust the individual polygon values? There is no easy question to this and needs further work.

2.4.9 What if not all selected polygons are sampled

It is difficult to predict in advance sample size requirements. Consequently, a useful strategy may to select more polygons than required; survey a sub-set; and if precision is insufficient; continue sampling. For example, a sample of 100 polygons may be initially selected, but only the first 80 surveyed. If precision is not sufficient, the remaining 20 may then be surveyed.

This causes no problems for the design. All that needs to be done is that the sampling weights be recomputed properly for the actual sample used as illustrated in Section 3.1. Also, the polygons to be surveyed should be selected at random from the list, i.e. don't just select the easy polygons to survey!

2.4.10 How many points should be selected within each polygon?

Under the current OS procedure, the number of points that is surveyed within each polygon is proportional to the area of the polygon. The PPSWR procedure does not require that this. Cochran (1977) discusses the allocation of sample in multi-stage designs - the optimal allocation requires information on the variability among points within polygons - this may be obtainable from past surveys. In many studies, most variation occurs at the primary stage and consequently, it is suspected that it is advantageous to select additional polygons with a single point per polygon rather than surveying multiple points within polygons. This should be verified empirically if possible.

3. Estimation

3.1 Estimating the total for each sampled polygon.

Regardless of the method used to estimate the overall population total, the first step is to obtain estimate of the total value of the variable in the selected polygons. This is obtained from the data collected in the ground survey. Regardless of the number of ground points measured in a specific polygon, the only information needed is the estimate of the total value for that polygon, denoted $\hat{Y}_i$. Because the polygons were selected with replacement, the way in which $\hat{Y}_i$ can be obtained is quite flexible. The design, the number of grid points, and the measurement methods used at each grid point can vary among the sampled polygons to adapt to different local conditions. For example, in one polygon, grid points may be selected using a simple random sample while in another polygon grid points may be selected using a systematic sample. In one polygon, local conditions may be difficult and only a single grid point can be measured while in another polygon, local conditions are easy and several grid points can be measured.

Regardless of the design within a polygon, of the number of points measured within a polygon, or of the way in which measurements are take, the important point is that the estimate $\hat{Y}_i$ be unbiased for the true polygon total Y_i. For example, measurements on a per hectare basis, such as volume/ha, need to be converted to total volume for the polygon by multiplying the volume/ha by the area of the polygon.

Consequently, the rest of this report will only deal with estimating TOTALS; the conversion of the total estimate and its se to a per-hectare or pre-tree basis is straightforward.

An interesting artifact of the with-replacement sampling design for selecting the polygons is that the individual ground points never appear to be used directly either in the

estimate of the population total, nor in the estimated variance. For example, if multiple grid points were selected within a polygon, then the individual measurements at each ground point explicitly appear neither in the formula for the estimate nor the formula for the estimated variance, i.e. an estimated polygon total based on a single grid point or 25 grid points is treated in exactly the same way and are given the same "weights". This is somewhat surprising as it appears that information is being ignored. Intuition would indicate to us that taking 25 measurements from polygons should give less variable results than taking only 1 measurement per polygon. However, all the information from all the grid points is being used – in the case of multiple grid points within a polygon, the estimate $\hat{Y}_i$ is less variable than an estimate based on a single grid points, and hence, the overall estimate is less variable and more precise. This situation is analogous to that encountered in the analysis of experimental designs where sub-sampling is present – the analysis proceeds using the averages over the sub-samples ignoring the individual sub-sampled values.

Similarly, when the various formulae for the variance of the estimator commonly found in text books are examined, it will be found that the formulae for the true (unknown) variance include terms for the second lower stage variabilities (e.g. among the ground points), but the estimated variances do not. Again, information is not being discarded – what happens is the estimated $\hat{Y}_i$ includes both polygon-to-polygon variation and all lower stage variation as well.

The implications of these results is that decisions about the number of ground points measured per polygon do have impacts on the final precision. The number of ground points measured per polygon also have direct cost implications. In the current procedures, the number of ground points measured per polygon is fixed by the total sample size and the area of the polygon. However, the PPSWR design gives additional flexibility in the allocation of resources between the number of polygons selected and the number of ground points measured per polygon - this could be used to further improve

precision for a given cost. Cochran (1977) has several sections on the allocation of resources between sampling at different stages of a multi-stage design.

Finally, it should be emphasized that the formulae for both the estimator and especially the estimated standard error of the estimator are tightly tied to the sampling procedure used to select the sample. If a different design from PPSWR is used, then the results of this paper are not applicable.

3.2 Estimates that use only the ground information

3.2.1 The Inflation Estimator

The simplest estimate of the population total uses only the estimated polygon totals from the selected polygons – the Phase I information is not used. This estimator is often called an **expansion** or **inflation** estimator because the estimated polygon totals are expanded or inflated to estimate the population total.

This is first done independently for each stratum in the population. For stratum h from which n_h polygons were sampled, the estimated population total for the stratum is computed using the Hansen-Hurwitz estimator as:

$$\hat{Y}_{h,\mathrm{inf}} = \frac{1}{n_h}\sum_{i=1}^{n_h}\frac{\hat{Y}_i}{p_i}$$

and its estimated variance (the se^2) is found as:

$$v\left(\hat{Y}_{h,\mathrm{inf}}\right) = \frac{1}{n_h}\frac{\sum_{i=1}^{n_h}\left(\frac{\hat{Y}_i}{p_i} - \hat{Y}_{h,\mathrm{inf}}\right)^2}{n_h - 1}$$

The grand total (over all strata) and the estimated variance of the grand total are found as:

$$\hat{Y}_{*,\text{inf}} = \sum_{h=1}^{H} \hat{Y}_{h,\text{inf}}$$

$$v\left(\hat{Y}_{*,\text{inf}}\right) = \sum_{h=1}^{H} v\left(\hat{Y}_{h,\text{inf}}\right)$$

Note that the variance of the grand total is found by adding the individual stratum variances and not by simply adding the individual stratum estimated standard errors (i.e. the se^2 terms must be added rather than simply summing the *se* terms.).

Once the total is obtained either at the stratum level or the inventory unit level, it can be converted to a "per hectare" basis by simply dividing it by the area of the stratum or inventory units. The standard error and confidence limits for the "per hectare" measurement is found by also dividing the standard error or confidence limits for the total, by the total area. When applied at the stratum level, this leads to:

$$\hat{\bar{Y}}_{h,\text{inf}} = \frac{1}{A_h}\hat{Y}_{h,\text{inf}} = \frac{1}{A_h}\frac{1}{n_h}\sum_{i=1}^{n_h}\frac{\hat{Y}_i}{p_i}$$

$$v\left(\hat{\bar{Y}}_{h,\text{inf}}\right) = \frac{1}{A_h^2}v\left(\hat{Y}_{h,\text{inf}}\right) = \frac{1}{A_h^2}\frac{1}{n_h}\frac{\sum_{i=1}^{n_h}\left(\frac{\hat{Y}_i}{p_i} - \hat{Y}_{h,\text{inf}}\right)^2}{n_h - 1}$$

where A_h is the area of stratum h. Combining over all strata, this gives:

$$\hat{\bar{Y}}_{*,\text{inf}} = \frac{1}{A}\hat{Y}_{*,\text{inf}} = \frac{1}{A}\sum_{h=1}^{H}\hat{Y}_{h,\text{inf}} = \frac{1}{A}\sum_{h=1}^{H}A_h\hat{\bar{Y}}_{h,\text{inf}} = \sum_{h=1}^{H}W_h\hat{\bar{Y}}_{h,\text{inf}}$$

$$v\left(\hat{\bar{Y}}_{*,\text{inf}}\right) = \frac{1}{A^2}v\left(\hat{Y}_{*,\text{inf}}\right) = \frac{1}{A^2}\sum_{h=1}^{H}v\left(\hat{Y}_{h,\text{inf}}\right) = \frac{1}{A^2}\sum_{h=1}^{H}A_h^2 v\left(\hat{\bar{Y}}_{h,\text{inf}}\right) = \sum_{h=1}^{H}W_h^2 v\left(\hat{\bar{Y}}_{h,\text{inf}}\right)$$

where A is the total area of the inventory unit, and W_h is the relative size of stratum h. This gives the standard result that the mean for the entire inventory unit is a weighted average of the stratum means.

A similar conversion from an estimate of a total to an estimate of a mean can be done for any estimator or sampling design.

An alternate estimator for the population total is the Horvitz-Thompson estimator:

$$\hat{Y}_{h,inf,HT} = \sum_{i=1}^{n_h} \frac{\hat{Y}_i}{\pi_i}$$

where π_i is the probability of selection in the entire sample (rather than at each draw) and for the PPSWR design is found as:

$$\pi_i = 1 - (1 - p_i)^{n_h}$$

i.e. the complement of being missed on every draw within the stratum. The Horvitz-Thompson estimator can be used for any sampling design - in most designs, the most difficult part is the computation of the inclusion probabilities for the estimate, and the computation of the joint inclusion probabilities (the probability that unit *i* and unit *j* will both be in the sample) for the variance estimate. Given the relatively simple form of the Hansen-Hurwitz estimators, there is no great advantage to using the Horvitz-Thompson estimator.

Before continuing, it is worthwhile explore the arguments often used that because individual grid points were selected so that every hectare in the inventory unit has the same probability of selection, the appropriate estimator is computed as simple average of the ground values. For consistency with the above result, assume that polygons within stratum *h* were selected with probability proportional to area and that ground points within each polygon were selected using a simple random sample, i.e every point had an equal chance of being selected. Let y_i represent the measurement at a ground location in polygon *i*, a_i represent the area of polygon *i* and A_h represent the area of all polygons in stratum *h*. Then the probability of selecting a one-hectare unit on each sample:

$$\text{prob any hectare selected} = \frac{a_i}{A_h} \times \frac{1}{a_i} = \frac{1}{A_h}$$

i.e. every hectare in the inventory unit's stratum has the same probability of selection and is given an equal weight. Consequently, the estimated stratum total can be computed by inflating the simple mean of all the data points:

$$\hat{Y}_{h,simple} = A_h \times \frac{\sum_{i=1}^{n_h} y_i}{n_h}$$

But, by rearranging terms, and noting that $a_i y_i = \hat{Y}_i$

$$\hat{Y}_{h,simple} = A_h \times \frac{\sum_{i=1}^{n_h} y_i}{n_h}$$

$$= \frac{1}{n_h} \sum_{i=1}^{n_h} A_h \frac{a_i y_i}{a_i}$$

$$= \frac{1}{n_h} \sum_{i=1}^{n_h} \frac{\hat{Y}_i}{a_i / A_h}$$

$$= \frac{1}{n_h} \sum_{i=1}^{n_h} \frac{\hat{Y}_i}{p_i} = \hat{Y}_{h,unf}$$

i..e. the two estimators are algebraically identical **assuming that the sampling plan proceeded without problems.** This was also noted by Stauffer (1995, p. 13). The advantage to expressing the estimator in the initial form is that there are **NO CHANGES** to the estimating equations even if every hectare is not given an equal chance of being selected. For example, under the equal-probability-for-each-hectare scheme, large polygons must have a larger number of ground points selected (on average) and no deviations within polygons are allowed, i.e. if a point cannot be sampled, great care must be exercised to choose another point within the same polygon. Under this more flexible scheme, the number of points sampled within each polygon can be chosen independently of the polygon size and there are no problems if the number of points actually sampled within each polygon differs from the theoretical specification. Consequently, there is no advantage to restricting the sample design so that hectares in different polygons have an equal chance of being selected. For example, it would be possible to have two equal size polygons and decide to allocate a single ground point to one polygon and several grid points to the second polygon. Note that within each specific polygon, each grid point normally will have an equal probability of selection compared to another grid point within the same polygon.

However, even though the two estimators may be algebraically identical when all assumptions for equal-hectare sampling are met, the theoretical precision is NOT computed as if each hectare were selected using a simple random sample because the

actual sampling design is not a simple random sample, but rather is a two stage design. In practice, Stauffer (1995) argues "Each point on the ground has an equal probability of being sampled, and, although the sampling is not SRS... provide justification for this [the SRS] estimator [of precision]" on the ground that the previous method of first sorting polygons by categories and by polygon area and then using a systematic sample to select ground points will lead to "the SRS estimator for the variance, though biased, will conservatively overestimate the variance". However, this argument breaks down if the design does not provide for each hectare to have an equal probability of selection. More to the point, why use an estimated variance that may be biased when the appropriate estimator for the variance is available and hardly more complicated to compute than the (invalid) SRS formula. Fortuitous coincidence is not a valid argument for using the SRS estimators.

This same arguments can be used for the ratio, regression, and geometric mean regression estimators in later sections, i.e. (1) the given formula have an algebraically identical simpler form if every hectare has an equal probability of selection but the simpler form are not robust to violations of the assumptions, and (2) why use a formula for the estimated precision based on an inappropriate design when the correct formula for this design are readily available and are automatically computed by most computer packages?

There are two possible methods of obtaining domain estimates. In the first method, the polygon value $\hat{Y}_i$ is replaced by a 0 if the polygon does not belong to the domain of interest and not changed for polygons belonging to the domain, i.e. create a new variable:

$$\hat{Y}_i^* = \begin{cases} \hat{Y}_i \text{ if polygon } i \text{ belongs to the domain} \\ 0 \text{ if polygon } i \text{ does not belong to the domain} \end{cases}$$

Then the above equations are used directly on the new variable for the entire sample to estimate the domain total, i.e.

$$\hat{Y}_{h,\text{inf},domain,\text{method 1}} = \frac{1}{n_h}\sum_{i=1}^{n_h}\frac{\hat{Y}_i^*}{p_i}$$

However, to estimate the domain mean, the domain total must be divided by the area in the population only in the domain – not the entire area of the inventory unit. For example, suppose that a domain of interest was polygons with Douglas fir as a leading species. All polygons where Douglas Fir was not the leading species would have the volume replaced by a zero. The estimated total is then the total volume for all such polygons, but to convert this to a per-hectare basis, the area of the polygons in the inventory units with Douglas Fir as a leading species would have to be known. Notice that the total number of polygons in each stratum does not need to be known. In some cases, this is difficult to determine from the information in the frame. All that is required is the number in the sample which can be determined from the ground measurements.

The second method of domain estimation estimates the average value per polygon in the domain and then multiplies this by the actual number of polygons in the domain. This is mathematically equivalent to a ratio estimator based on the entire sample, i.e for each stratum the domain estimator is computed as:

$$\hat{Y}_{h,\text{inf},domain,method\ 2} = \frac{estimated\ \ total\ \ of\ \ values\ \ in\ \ domain}{estimated\ \ number\ \ of\ \ polygons\ \ in\ \ domain} \times Number\ \ of\ \ polygons\ \ in\ \ domain$$

$$= \frac{\sum_{i=1}^{n_h}\frac{\hat{Y}_i^*}{p_i}}{\sum_{i=1}^{n_h}\frac{I_i^*}{p_i}} \times Number\ \ of\ \ polygons\ \ in\ \ domain$$

where

$$\hat{Y}_i^* = \begin{cases} \hat{Y}_i \text{ if the polygon belongs to the domain} \\ 0 \text{ if the polygon does not belong to the domain} \end{cases}$$

$$I_i^* = \begin{cases} 1 \text{ if the polygon belongs to the domain} \\ 0 \text{ if the polygon does not belong to the domain} \end{cases}$$

The estimated standard error is determined using the methods found in Section 3.3.1. Despite its 'strange' appearance, Method 2 is nothing more than simply finding the sample

average of units that only belong to the domain and then multiplying by the total number of polygons belonging to the domain.

The first method must be used if the domain information (specifically the number of polygons of the domain in the entire unit) is not available in the frame. For example, suppose the domain is defined as under attack by insects. This is not likely available in the frame (based on the Phase I information) and is only available from ground samples. The second method can be used in situations where the frame does contain information that allows population units (and elements of the sample) to be classified into the domain..

Both methods are illustrated in the detailed example later.

3.3 Estimates that use the relationship between the Phase I and Phase II values.

Precision of the estimates can often be improved by using the relationship between the Phase I and Phase II values. There are three common ways of using this relationship:

- Ratio estimators that assume a linear relationship through 0 between the two variables
- Regression estimators that assume a linear relationship not through 0
- Geometric Regression estimators that assume a linear relationship but allows both Phase I and Phase II values to include "errors" in measurement.

To choose among these estimators, the analyst should first make a plot of the relationship between the Phase I and Phase II values for the selected polygons and determine if the relationship is linear and if it passes through the origin. In addition, the analyst should investigate any apparent outliers for coding or other errors.

In some circumstances, it is possible to use more than one variable in the adjustment process. Such multivariate ratio, regression, or geometric mean regression estimators are beyond the scope of this report, but are discussed by Cochran (1977).

Note that in the VRI, the sampling design provides information at two different levels in the design. The auxiliary information (the Phase I variables) is available at the polygon level while the elemental unit information (the Phase II variables) is available at the ground point level. Sarndal et al (1992, Chapter 8) discusses these and other situations (e.g. auxiliary information available at elemental levels etc).

3.3.1 Ratio Estimators

In the ratio estimator, the estimated ratio between the Phase I and II total within the sample is used to adjust the Phase I population total. This is known as model-assisted sampling – the presumed relationship between the Phase I and Phase II totals is used to improve the estimation process (Särndal et al., 1992).

There are two forms using ratio estimators – these are commonly called separate and combined ratio estimators.

In the separate ratio estimator, separate ratio estimates of the population total are formed for each stratum, and then the estimated stratum total are added together. One common model assumes that in each stratum the variance in the response is proportional to the Phase I total. Under this model, the estimated total for each stratum is found as:

$$\hat{Y}_{h,ratio} = \hat{R}_h X_h = \frac{\hat{Y}_{h,inf}}{\hat{X}_{h,inf}} X_h = \frac{\frac{1}{n_h}\sum_{i=1}^{n_h}\frac{\hat{Y}_i}{p_i}}{\frac{1}{n_h}\sum_{i=1}^{n_h}\frac{X_i}{p_i}} X_h = \hat{Y}_{h,inf} + \hat{R}_h\left(X_h - \hat{X}_{h,inf}\right)$$

There are several estimates of the variance of the estimator (Särndal et al. 1992; Section 7.2). One simple estimate is:

$$v\left(\hat{Y}_{h,ratio}\right) = \frac{1}{n_h}\frac{\sum_{i=1}^{n_h}\left(\frac{e_i}{p_i} - \frac{1}{n_h}\sum_{i=1}^{n_h}\frac{e_i}{p_i}\right)^2}{n_h - 1} = \frac{1}{n_h}\frac{\sum_{i=1}^{n_h}\left(\frac{e_i}{p_i}\right)^2}{n_h - 1} \text{ where } e_i = \hat{Y}_i - \hat{R}_h X_i$$

The grand total and estimated variance of the grand total are found as before:

$$\hat{Y}_{\bullet,ratio} = \sum_{h=1}^{H}\hat{Y}_{h,ratio}$$

$$v\left(\hat{Y}_{\bullet,ratio}\right) = \sum_{h=1}^{H} v\left(\hat{Y}_{h,ratio}\right)$$

Note that the variance of the grand total is found by adding the individual stratum variances and not by simple adding the individual stratum estimated standard errors (i.e. the se^2 terms must be added).

The separate ratio estimator requires information on the population total of the Phase I values for each stratum – something that is usually available in the VRI. In addition, individual stratum estimates are also obtained – again this is usually of interest.

An alternate ratio estimator is the **combined** ratio estimator formed as;

$$\hat{Y}_{\bullet,comb\ ratio} = \frac{\sum_{h=1}^{H}\hat{Y}_{h,\inf}}{\sum_{h=1}^{H}\hat{X}_{h,\inf}} X$$

i.e. a single ratio is formed from the totals overall all strata and then multiplied by the overall total of Phase I variable. The combined ratio estimator does not need the individual stratum totals for the auxiliary variable, and assumes that the ratio between the two variables is the same in all strata. As Cochran (1977) notes, unless the ratios within each stratum are comparable, the separate ratio estimator is likely more precise. Furthermore, the separate ratio estimator does provide stratum specific estimates. For these two reasons, the combined ratio estimator is not recommended but with limited

sample sized some strata ratios may be highly erratic. Some hybrid (shrinkage) method combining the virtues of the separate and combined ratio estimator may give better performance but is beyond the scope of this report.

The ratio estimators assume that the relationship between the response and auxiliary variable is linear through the origin. It performs well, if the variability in the response variable also increases with the auxiliary variable. The guidelines of Stauffer (1995, p.29-32) on model assessment and fit are also important.

There are also two methods of domain estimation similar to those used in the inflation estimator. In Method 1, the response variable is replaced by the value of 0 for polygons not in the domain before computing the ratio. In Method 2, a separate ratio is determined for the domain of interest based on units belonging solely to the domain and this is multiplied by the domain total Phase I variable. The conditions under which Method 1 and Method 2 can be used are similar to the previous section. In addition, Method 1, although leading to a proper estimate of the total, is somewhat unsatisfactory as the estimated ratio is no longer the intuitive ratio between the Phase 2 and Phase 1 variable, but some sort of average ratio that includes many zeroes for the Phase 2 variable. It is difficult to interpret, has a much larger sampling variation compared to Method 2, and cannot be used to adjust individual polygon values. Consequently, this method is not to preferred except in situations where it must be used (i.e. domain classification depends on data from Phase II and cannot be done based on information in the frame).

Note that with the advent of modern computer packages that **properly use survey data in regression type models**, many of the above methods are just different models fit to the same data. For example, the following table illustrates the correspondence between the above estimators, models that can be fit, and a "SAS-type specification" in a generalized modelling framework.

Correspondence between ratio estimators and SAS syntax		
Estimator	Statistical model	SAS-type syntax
Separate ratio estimator	$E[Y_{hj}] = R_h X_{hj}$	Class stratum; Y =X stratum*X / noint;
Combined ratio estimator	$E[Y_{hj}] = RX_{hj}$	Class stratum; Y = X / noint;
Method 2 Domain estimator with separate slopes in each stratum h=stratum; d=domain	$E[Y_{hj}] = R_{hd} X_{hj}$ R=ratio	Class stratum domain; Y=X stratum*X domain(stratum)*X / noint;

Notice the strong correspondence between the syntax used here and that used for the analysis of covariance where slopes may be equal or unequal for the factor levels.

An important task when using a computer package to fit survey data is to verify the model being fit to the data – in particular the modelled variance function. Different assumptions about the variance function (i.e. is it proportional to X or is it constant) will lead to different estimates of the ratio, different estimates of the population total, and different variance estimates.

Modern packages also provide model testing statistics to help the analyst choose the best fitting model. Once the best fitting model is chosen, the estimated value of the ratio is multiplied by the appropriate total from Phase 1. With these model fitting capabilities now available, many other models are easily fit that don't have explicit formulae, e.g. a domain estimator with the same slope for the domain in all of the strata, or multiple ratio variables - these are beyond the scope of this report. Note that ordinary regression fitting routines do not account for the survey design and should not be used in the model fitting procedure.

3.3.2 Regression Estimators

In regression estimators, the relationship between the Phase I and Phase II polygon total is again assumed to be linear, but not necessarily passing through the origin. The basic concept is to estimate the regression line between the two phases and use the regression line to adjust the inflation estimator.

Before applying this estimator, plots should be made to investigate if the relationship is linear and the pattern of variability.

As in the ratio estimator, there are two possible forms of the regression estimator.

The form of the regression estimators will depend upon the assumed population model (Sarndal et al, 1992; Chapter 7). Under a model where there is a linear relationship between the Phase I and Phase II values with a constant variance, the estimated intercept and slope for each stratum for the separate regression estimator, are derived following Särndal et al (1992, p.230, Remark 6.4.4) as:

$$a_h = \frac{\hat{Y}_{h,inf} - b_h \hat{X}_{h,inf}}{\frac{1}{n_h}\sum_{i=1}^{n_h}\frac{1}{p_i}}$$

$$b_h = \frac{\frac{1}{n_h}\sum_{i=1}^{n_h}\frac{X_i\hat{Y}_i}{p_i} - \frac{\hat{X}_{h,inf}\hat{Y}_{h,inf}}{\frac{1}{n_h}\sum_{i=1}^{n_h}\frac{1}{p_i}}}{\frac{1}{n_h}\sum_{i=1}^{n_h}\frac{X_i^2}{p_i} - \frac{\hat{X}_{h,inf}^2}{\frac{1}{n_h}\sum_{i=1}^{n_h}\frac{1}{p_i}}} = \frac{\frac{1}{n_h}\sum_{i}^{n_h}\frac{\left(X_i - \tilde{X}_h\right)\left(Y_i - \tilde{Y}_h\right)}{p_i}}{\frac{1}{n_h}\sum_{i}^{n_h}\frac{\left(X_i - \tilde{X}_h\right)^2}{p_i}}$$

where $\tilde{X}_h = \hat{X}_{h,inf} \Big/ \frac{1}{n_h}\sum_{i=1}^{n_h}\frac{1}{p_i}$, $\tilde{Y}_h = \hat{Y}_{h,inf} \Big/ \frac{1}{n_h}\sum_{i=1}^{n_h}\frac{1}{p_i}$.

The estimated slope and intercept are used to predict the total for each polygon and the estimated population total for stratum h is found by summing all predicted values:

$$\hat{Y}_{h,reg} = \sum_{i=1}^{N_h} (a_h + b_h X_i) = N_h a_h + b_h X_h \approx \hat{Y}_{h,inf} + b_h (X_h - \hat{X}_{h,inf})$$

There are several variance estimators (Särndal et al.1992) and one simple estimator has the form:

$$v\left(\hat{Y}_{h,reg}\right) = \frac{1}{n_h} \frac{\sum_{i=1}^{n_h} \left(\frac{e_i}{p_i} - \frac{1}{n_h} \sum_{i=1}^{n_h} \frac{e_i}{p_i} \right)^2}{n_h - 1} = \frac{1}{n_h} \frac{\sum_{i=1}^{n_h} \left(\frac{e_i}{p_i} \right)^2}{n_h - 1} \text{ where } e_i = \hat{Y}_i - (a_h + b_h X_i)$$

The grand total and estimated variance of the grand total are found as before:

$$\hat{Y}_{*,reg} = \sum_{h=1}^{H} \hat{Y}_{h,reg}$$

$$v\left(\hat{Y}_{*,reg}\right) = \sum_{h=1}^{H} v\left(\hat{Y}_{h,reg}\right)$$

Note that the variance of the grand total is found by adding the individual stratum variances and not by simple adding the individual stratum estimated standard errors (i.e. the se^2 terms must be added).

A combined regression estimators can also be formed, but it, like the combined ratio estimator assumes that the relationship between the two variables is equal across strata. For the same reasons as for the combined ratio estimator, it is unlikely that a combined regression estimator will be of much use.

The regression estimator performs well when the relationship between the two variables is linear and that the variation is relatively constant over the entire regression line.

Both methods of domain estimation can also be used. Again, in Method 1, the response variable is set to 0 for polygons not part of the domain and the above equations used without changes. As for the ratio estimator, the estimated slope is not readily interpreted and the estimate is likely to have high variance. In Method 2, a separate slope and intercept are fit using only polygon that belong to the domain.

These models can also be cast into a general framework. For example, the following table illustrates the correspondence between the above estimators, models that can be fit, and a "SAS-type specification":

Correspondence between regression models and SAS syntax		
Estimator	Statistical model	SAS-type syntax
Separate ratio estimator	$E[Y_{hj}] = \alpha_h + \beta_h X_{hj}$	Class stratum; Y =stratum X stratum*X ;
Combined ratio estimator	$E[Y_{hj}] = \alpha + \beta X_{hj}$	Class stratum; Y = X ;
Method 2 Domain estimator with separate slopes in each stratum	$E[Y_{hj}] = \alpha_{hd} + \beta_{hd} X_{hj}$	Class stratum domain; Y=domain(stratum) X domain(stratum)*X;
h=stratum; d=domain	α=intercept; β=slope	

Once the best fitting model is chosen, the estimated line is used with the Phase 1 totals to estimate the Phase 2 total. Many other models are easily fit that don't have explicit formulae, e.g. a domain estimator with the same slope for the domain in all of the strata or multiple regression variables; these are beyond the scope of this report. Note that ordinary regression fitting routines that do not account for the survey design should not be used in the model fitting procedure as they fail to account for the ways in which the sample was selected. It is possible to extend this method to more than one predictor variable, but there are no explicit closed-form formulae.

3.3.3 Geometric Mean Regression Estimators

It is usually assumed in linear regression that the X variable is measured without error and that all of the variation occurs in the response variable. In the VRI project, this is clearly not the case. Both the Phase I and Phase II measurements are only estimates of the

true polygon value. In cases where both X and Y are subject to variation around the true value, estimates of the slope for the relationship between X and Y are known to be biased downwards (Berkson 1950), and cases are often made for "error in variable methods", of which the geometric mean regression is one example.

However, presence of "error" in X does not automatically imply that estimates of the slope are biased. Berkson (1950) showed that if there is no correlation between the intended X value and the apparent response error, then the ordinary regression estimates of the slope are unbiased. For example, when applying a herbicide, the concentration is specified in advance. However, because of internal variability in the applicator, the actual concentration differs from the nominal concentration. The response depends upon the actual concentration plus a random factor. Here the apparent response error in Y is uncorrelated with the nominal concentration and so the ordinary regression estimates remains unbiased.

An example of where bias in the regression estimate would occur, would be a case where a watering device delivers an amount of water and measuring devices are set in the field to measure the amount delivered. Yield of the crop is the response variable. Here the measuring devices in the field measure the actual amount delivered plus measurement error. Now there is a non-zero correlation between the intended X value (the measurement taken of the amount of water delivered) and the apparent response error at each actual amount of water delivered and the estimated slope will have a negative bias. [Places where the amount of water was higher than nominal (a positive measurement error) will likely have plant growth greater than the average for that nominal X value (a positive response error)].

The magnitude of the bias is related to the magnitude of the measurement error relative to the variation along the X axis. For example, if the standard deviation of the measurement error is 10% of the standard deviation of the values along the X axis, the slope is biased by a factor of about 1%. (Angleton and Bonham, 1995). Consequently,

unless the error in the X values is an appreciable fraction of the variability in the X values, any biases are certainly negligible.

The geometric mean regression is often uncritically recommended for error-in-variable problems. However, many users are unaware of the potential inconsistencies in the estimator. First, if no assumptions are made about the relative sizes of the two measurement error, no estimators perform well as the problem has more parameters than can be estimated from the data. Second, the geometric mean regression is also biased unless certain assumptions about the ratio of the two error variances to certain statistics in the data unlikely to be satisfied in practice occur. Third, if the ratio of the error variances is known (e.g. it seems plausible that the error variances could be equal for both variables), then better methods are available. Draper and Smith (1997), and Riggs et al (1978) have a complete discussion about these points.

Given these caveats, the estimated slope and intercept for stratum *h* are found as:

$$b_{h,gmr} = \sqrt{\frac{b_{h,Y\,on\,X}}{b_{h,X\,on\,Y}}} \times sign\ of\ slope \text{ and}$$

$$a_{h,gmr} = \bar{\hat{Y}}_{h,inf} - b_{h,gmr}\bar{\hat{X}}_{h,inf}$$

where $\bar{\hat{Y}}_{h,inf}$ *and* $\bar{\hat{X}}_{h,inf}$ are the mean per polygons based on the inflation estimators.

Then, as in the regression estimator, the estimated slope and intercept are used to predict the total for each polygon and the estimated population total for stratum *h* is found by summing all predicted values:

$$\hat{Y}_{h,gmr} = \sum_{i=1}^{N_h}\left(a_{h,gmr} + b_{h,gmr}X_i\right) = N_h a_{h,gmr} + b_{h,gmr}X_h = \hat{Y}_{h,inf} + b_{h,gmr}(X_h - \hat{X}_{h,inf})$$

with a simple variance estimator of:

$$v\left(\hat{Y}_{h,gmr}\right) = \frac{1}{n_h}\frac{\sum_{i=1}^{n_h}\left(\frac{e_i}{p_i} - \frac{1}{n_h}\sum_{i=1}^{n_h}\frac{e_i}{p_i}\right)^2}{n_h - 1} = \frac{1}{n_h}\frac{\sum_{i=1}^{n_h}\left(\frac{e_i}{p_i}\right)^2}{n_h - 1} \text{ where } e_i = \hat{Y}_i - \left(a_{h,gmr} + b_{h,gmr}X_i\right)$$

The grand total and estimated variance of the grand total are found as before:

$$\hat{Y}_{\cdot,gmr} = \sum_{h=1}^{H} \hat{Y}_{h,gmr}$$

$$v\left(\hat{Y}_{\cdot,gmr}\right) = \sum_{h=1}^{H} v\left(\hat{Y}_{h,gmr}\right)$$

A combined geometric mean regression is also unlikely to be useful in practice.

Domain estimation proceeds in a similar fashion as in the regression estimators.

Each case must be assessed individually, but it seems unlikely that the geometric mean regression will provide improvements in the estimation process unless the errors in both measurements are considerable and comparable to the spread of the X values observed in the sample. As well, from a design-based view, the fact that the observed Phase I value is not equal to the actual value in the polygon is immaterial - the sampling process still guarantees that the estimates are unbiased in ordinary ratio or regression estimates as long as there is some relationship between the phase I and phase II measurements. The real problems will occur when individual polygon values are adjusted or inverse predictions are make (i.e. estimate the Phase I values from the Phase II values). In both cases, it is unlikely that the uncertainty in the X measurements has been included.

More complex multiple regression error-in-variable models can also be fit, but these are extremely complex and beyond the scope of this report.

3.4 Confidence intervals

Regardless of the method used to compute the estimate and its estimated standard error, an approximate confidence interval can be found by recourse to the central limit theorem which states that regardless of the population structure, estimates based on means typically follow a normal distribution in large sample. Hence an approximate confidence interval is found as:

$$estimate \pm z_{\alpha} se$$

where z_α is the appropriate percentage point of a normal distribution. For example, for 95% confidence intervals, $z_\alpha = 1.96$.

Caution should be exercised when sample sizes are "small" or responses have a large skewness, e.g.. a long right tail. In these circumstances, the above estimated confidence interval may not perform well and severely underestimate the actual confidence in the interval.

Unfortunately, there is no formal way to determine if the sample size is sufficiently large enough to ensure that the above confidence interval performs well. Stauffer (1995, p. 39) recommends that "minimum sample sizes n=30, and, preferably, n=50 be used for the strata, where separate regression estimates are required."

An alternative method to estimate the standard errors would be through the use of a bootstrap. In this procedure, the observed data are resampled with replacement to create a series of "new" samples; the estimates are obtained from this "new" sample; and the actual distribution of estimates over the set of "new" samples are used to determine the lower and upper confidence limits. Rao (1997, Section 3.1) and Sarndal et al (1992, Section 11.6) describes the construction of the bootstrap for complex survey design.

3.5 Adjusting individual polygon values

Stauffer (1995, p.34-38) summarizes many of the options available for adjusting individual polygons. In particular, once the appropriate ratio or regression model is fit, these are applied at the individual polygon level. However, his recommendations on the precision of the adjusted values must be adjusted to account for the survey design by replacing the formulae he gives with analogues using sampling weights as seen in previous sections.

It should be noted that there are two, possible distinct goals, involved in the inventory process. One goal may be good estimation of the population total while a second goal may be to adjust individual polygon totals as well as possible. Several methods may be used for a particular survey, as one method may not be good for both purposes. For example, if adjustments are to be made to several variables and the adjusted totals must all be consistent with each other and the estimated totals then a regression estimator working with a single variable at a time, may not lead to consistent estimators. Stauffer (1995, p.38) has a brief discussion of this problem.

Thrower (1998) also prepared a series of reports on the adjustment of individual polygons. The final report included a detailed example of the Boston Bar example.

4. Computer Software

Unlike the analysis of experimental design data, software for the analysis of survey data is not readily available, but several standalone packages are available ranging in price from free to several thousand dollars per year. A review of existing software and the reasons why specialized software is necessary for the analysis of survey data is provided by the Survey Methods Section of the American Statistical Association at:

http://fas-www.harvard.edu/~stats/survey-soft/survey-soft.html

Fortunately, SAS Version 8 now contains procedures for the analysis of survey data that are integrated into the SAS system. There are three procedures:

Proc SURVEYSELECT – to select samples
Proc SURVEYMEANS – for estimates using inflation methods
Proc SURVEYREG – for estimation using ratio and regression methods.

These new procedures in SAS do not use any new methodology compared to the other packages and have similar capabilities. The advantage of using SAS over a standalone package is that all the data management and analysis tools already present in SAS can be used in conjunction with these new procedures without having to learn a new package.

Consequently, this report will demonstrate the methods presented in the previous sections using the SAS system.

It is recommended that proper computer packages specially designed for the analysis of survey data (e.g. SAS V.8) be used to process the data rather than trying to implement the formulae in this report using, for example, a spreadsheet. The latter often leads to errors in transcribing the formulae and numerical problems for large datasets. A sample set of SAS programs was created to analyze the simulated example in the Appendix. These can serve as templates for the analysis of a real survey. Note that the estimated variances from SAS may differ slightly from those computed using the simple variance formula as SAS uses a more complex estimator.

5. Example

Eventually hyperlinks from report to the files given will be added. For now refer to

www.math.sfu.ca/~cschwarz/MOF

to access copies of the files.

5.1 The Phase I data

A simulated population was created based upon data from the Sunshine Coast provided by Dr. M Penner on behalf of the Ministry of Forests.

It consists of approximately 46,000 polygons covering about 800,000 ha. From the full Phase I data available, the following variables were extracted for use in this example: The codes used for a particular variable and their meaning are available from an on-line data dictionary at MoF (http://www.for.gov.bc.ca/resinv/reports/rdd/search/rddseaan.htm)

Phase I variables selected from the SunShine Coast Dataset	
Variable	Meaning
map_no	the map sheet number that contains the polygon
polygon	the polygon number on the map sheet. Both the map_no and polygon value are needed to uniquely identify a particular polygon.
polyarea	The area of the polygon (ha)
esa1_cd	Environmentally sensitive area code. Refer to the data dictionary for details of code values. A blank value implies a polygon is not in an evironmentally sensitive area.
npd_cd	Non-productive code. Refer to the data dictionary for details of the code values. A blank value implies a productive polygon.
sspcs1	Species code for leading species in polygon. Refer to data dictionary for details of code values.
sspcs2	Species code for secondary species in polygon.
sspcs3	Species code for third species in polygon.
vol11	Net volume per hectare of the leading commercial species at the primary utilization level. Net volume per hectare is determined as gross volume less decay, waste, and breakage. Refer to the data dictionary for additional details
vol12	Net volume per hectare for the secondary commercial species at the primary utilization level.
vol13	Net volume per hectare for the third commercial species at the primary utilization level.
bio_geo	Biogeographic climatic zone. Refer to the data dictionary for details of code values

Based upon the above variables, the following derived variables were created:

Derived variables for the Sunshine Coast Example	
esa	derived value. "yes" = ESA, "no" = not an ESA
phase1_volha	vol11 + vol12 + vol13
phase1_vol	phase1_volha * polyarea /1,000,000= total net yield from polygon The factor 1,000,000 is used simply to scale the results to a more manageable range. Note that this variable measures the total volume for the entire polygon rather than on a per-hectare basis.

The complete list of polygons extracted from the SunShine Coast is found in the file frame.dat.

The non-vegetated polygons (those with non-blank non-productive code or a blank leading species code) are of little interest in the VRI and hence will be deleted from the population of interest. The resulting population consists of about 36,000 polygons covering about 550,000 ha.

Summary statistics about the reduced frame are found in read.frame.lst. Of the vegetated polygons, about 80% of the polygons are in the cedar-western-hemlock biogeoclimatic zone..

5.2 Deciding upon strata, sample sizes, and allocation.

The VRI could be used as an ecologically based inventory. Because some tree species have particular ecosystem preferences, it is useful to stratify the population of vegetated polygon by leading species.

A tabulation of the Phase I data by leading species is found in optimal.allocation.lst. About 2/3 of the polygons have Douglas Fir (species code FD) or Hardwood (species codes starting with H) as leading species. To investigate the allocation

of samples among the strata, the first character of the species code will be used as an initial stratification variable.

Following Cochran (1977, p.253), the theoretical variance of the inflation estimator within each stratum *h* is:

$$V\left(\hat{Y}_{h,inf}\right) = \frac{1}{n_h}\sum_{i=1}^{N_h} p_i\left(\frac{Y_i}{p_i} - Y_h\right)^2$$

In order to investigate the effects of different allocations, the theoretical value above must first be known – this is clearly impossible as the Phase II values can never be known for the entire population. However, it seems reasonable to use the Phase I values as surrogates for the true population values. These will, of course, not be the theoretical variances – rather it is hoped that these results will provide guidance on what to expect in the actual survey. Consequently, the "variance" of an estimator within each stratum can be computed as:

$$Var_h = \sum_{i=1}^{N_h} p_i\left(\frac{Y_i}{p_i} - Y_h\right)^2$$

and then the total variance for a particular allocation of n_h to each stratum is found as:

$$Var_{total} = \sum_{h=1}^{H} \frac{Var_h}{n_h}$$

based on the Phase I value.

The Var_h values were computed for the Phase I volume per polygon using optimal.allocation.sas and are shown in optimal.allocation.lst. The effects of different allocation schemes can then be investigated easily using an Excel spreadsheet available with this report. This spreadsheet computes the theoretical variance for an equal allocation of the total sample over all strata, an allocation proportional to the total polygon area of each stratum, and an optimal allocation where the SOLVER feature of Excel is used to minimize the overall variance keeping the total sample size fixed. The results for a total sample size of 200 are as follows:

Effects of different allocations on the inflation estimator based on Phase 1 Data

Species	N polygons	Total area	Stratum Var	Equal Allocation n_h	Equal Allocation Var of Inflation est	Allocation Proportional to area n_h	Allocation Proportional to area Var of Inflation est	Optimal Allocation n_h	Optimal Allocation Var of Inflation est
A	377	3975.3	0.16	18.2	0.009	1.4	0.114	0.8	0.207
B	3104	52944.7	112.07	18.2	6.164	18.8	5.972	20.4	5.484
C	2194	33943.2	52.62	18.2	2.894	12.0	4.374	14.0	3.758
D	3137	43553.1	28.24	18.2	1.553	15.4	1.829	10.3	2.753
E	18	154.8	0.10	18.2	0.006	0.1	1.823	0.6	0.164
F	13451	212713.1	1805.39	18.2	99.296	75.4	23.948	82.0	22.012
H	11898	189442.8	1166.50	18.2	64.158	67.1	17.374	65.9	17.694
M	245	3450.2	0.13	18.2	0.007	1.2	0.106	0.7	0.187
P	894	15345.8	0.98	18.2	0.054	5.4	0.180	1.9	0.513
S	172	2936.6	0.39	18.2	0.021	1.0	0.375	1.2	0.324
Y	354	5846.8	1.26	18.2	0.069	2.1	0.608	2.2	0.582
	35844	564306.4		200.0		200.0		200.0	
		Overall Variance			174.231		56.703		53.677
		CV (%)			7%		4%		4%
Non-stratification									
All	35844	564306.4	18363.69	200	91.818				
	Total vol	198.3		CV	5%				

First, the variance of the total for the inflation estimator when no stratification is done is expected to be about 92 (i.e. a standard error of about 10) giving a cv of about 5%. Equal allocation (about 20 polygons selected from each stratum) leads to variance that is about double that of non-stratifying – this is a very poor choice. Allocating samples proportional to the total area of the polygons within the strata leads to about a 50% reduction in the variance; an optimal allocation only leads to a slight improvement. This is in agreement with Cochran (1977) who notes that fixed allocations are usually worse than proportional allocations which in turn are usually only marginally worse than optimal allocations.

If the level of precision is inadequate, then it is a simple matter to increase the total sample size until precision targets are reached. Note that the above exercise is only an approximation to what will happen in the actual survey as it is assumed that what happens with the Phase I information is a good surrogate to what will happen with the Phase II data.

From the above table, it is decided that three strata are appropriate; F,H, and others. The allocation of samples will be about 80:70:50 for the three strata, based on an approximate proportional allocation by total polygon area.

Summary statistics on the strata are presented in select.sample.sas.

5.3 Selecting the sample

As noted previously, polygons will be selected with a PPSWR design based on the polygon area. There are a number of polygons whose polygon area is recorded as 0, including a few productive polygons with what appear to be a fair amount of timber. These are listed in select.sample.lst and should be investigated further.

The SAS procedure PROC SUREVEYSELECT was used to select the polygons as shown in select.sample.sas.. One polygon was selected twice as shown in select.sample.lst. This polygon will be "ground sampled twice". The unique set of 199 polygons is shown in select.sample.lst.

5.4 Obtaining the Phase II data

Following Stauffer (1995, p.46) the Phase II ground readings for a polygon in the productive strata are generated as:

$$phase2_volha = 100 + .9 phase1_volha + error$$

where the error variability is assumed to normally distributed with a mean of 0 and a standard deviation of 25% of the Phase 1 volume/ha reading. If the generated Phase 2

reading was less than zero, it was replaced by its absolute value. [This is purely arbitrary and the reading could have been replaced by zeroes. All that matters is that for this example, some Phase II reading is available.]

Multiple samples within the same polygon will be given different readings.

Approximately 10% of the Phase II readings were randomly set to missing values to simulate the effect of missing data. Similar procedures as outlined in the rest of this section would be followed if polygons were added, i.e. simply adjust the sampling weights.

The code segment get.phase2.sas returns the Phase II readings for the selected polygons.

5.5 The Inflation estimates

The SAS procedure PROC SURVEYMEANS was used to estimate the total volume for each individual stratum and the entire inventory unit based only on the Phase II readings. The file inflation.est.sas illustrates the use of the procedure to estimate the total volume for the inventory unit.

As noted earlier, the variable of interest is the total volume from the Phase II survey for each polygon, i.e the Phase II volume/ha times the polygon area.

There were 23 polygons that had missing data for the Phase II readings. Most computer packages that deal with survey data, compute estimates based on sampling weights. The sampling weight for an observation is a number that represents the contribution of this observation to the estimation of the total. For example, the inflation estimator for a stratum total can be expressed as:

$$\hat{Y}_{h,\inf} = \frac{1}{n_h}\sum_{i=1}^{n_h}\frac{\hat{Y}_i}{p_i} = \sum_{i=1}^{n_h}\frac{\hat{Y}_i}{n_h p_i} = \sum_{i=1}^{n_h} w_{hi}\hat{Y}_i$$

where

$$w_{hi} = \frac{1}{n_h p_i} = \frac{Z}{n_h z_i}$$

is the sampling weight for the i^{th} observation in stratum h.

Many software packages automatically compute the sampling weights when they select a sample from the population. However, because these sampling weights would change if observations are dropped or added, the initial sampling weights cannot be used and the sampling weights must be recomputed when the estimation is done.

These weights are recomputed using the above equation, with the revised sample sizes after the polygons with missing values are deleted. The code fragment compute.samplingweights.sas illustrates how to recompute the sampling weights in the presence of missing data using a common non-response adjustment factor as suggested by Rao (1997, Section 4).

The SURVEYMEANS procedure was used twice – one to obtain the stratum specific estimates of the volume and the second time to obtain the estimate over all strata. The results are shown in inflation.est.lst and are summarized in Section 5.9.

Domain estimates were obtained for polygons that belong to the cedar-western-hemlock (CWH) biogeoclimatic zone. These were obtained using Method 1 and Method 2 and the estimates are presented in inflation.est.lst. The estimates appear to be "different" but are consistent with each other given the precision of each estimate.

5.6 The Ratio Estimates

It seems reasonable that there should be a relationship between the Phase 1 volume estimate and the Phase II volume estimate. This is exploited in computing a separate-ratio estimate as shown in the file ratio.est.sas

Plots of the Phase II volume against the Phase I volume showed a linear relationship that didn't quite pass through the origin, but should be close. These plots are found in ratio.est.lst and are summarized in Section 5.9.

The estimate (found in ratio.est.lst) is approximately equal to that from the inflation estimator, but the standard error is about 25% smaller because of the relationship between the two variables.

Domain estimates were obtained for polygons in the cedar-western-hemlock (CWH) biogeoclimatic zone using Method 1 and Method 2 and the estimates are presented in ratio.est.lst. These estimates are similar to those found using an inflation estimator, but the c.v.'s are slightly larger for Method 1. This is not surprising as replacing the Phase II volume with 0 surely destroys the linearity relationship between the two variables. The precision for Method 2 is better compared to the inflation domain estimator; the strong relationship between the two variables improves the estimation.

5.7 The Regression Estimates

Regression estimates were computed using the code in regression.est.sas.

The estimates (found in regression.est.lst) is approximately equal to that in the previous sections, but the standard error is slightly smaller than the ratio estimator, but of little practical importance. The results are summarized in Section 5.9

Domain estimates are again computed using Method 1 and Method 2. Method 1 is not recommended in this case because the domain total are known in advance.

5.8 Geometric Mean Regression Estimators

Regression estimates were computed using the code in geomean.est.sas.

The estimates (found in geomean.est.lst) is approximately equal to that in the previous sections. The results are summarized in Section 5.9

Domain estimates are again computed using Method 1 and Method 2. Method 1 is not recommended in this case because the domain total are known in advance.

5.9 Summary of the estimates

Table 5.9 presents a summary of the estimates computed for both the entire inventory unit and the cedar-western-hemlock biogeoclimatic zone. The volume/ha estimates are derived by dividing the estimated total volume (and its estimated se) by the area of the stratum or unit respectively.

Table 5.9 Summary of the estimates for the total volume over the inventory unit computed using the various methods

	Total volume (millions of m^3) for the entire inventory unit		Volume per hectare (m^3/ha) for the entire inventory unit.[1]		Total volume (millions of m^3) for the CWH Domain[2] computed using domain estimator Method 1		Total volume (millions of m^3) for the CWH Domain[2] computed using domain estimator Method 2	
Stratum	Estimated Total	Est SE	Estimated volume/ha	Est SE	Estimated Total	Est SE	Estimated Total	Est SE
Inflation Estimator								
F	90.95	6.11	427.5	28.7	79.0	7.0	89.0	13.9
H	78.58	5.29	414.8	27.9	51.3	6.2	50.7	7.6
Other	53.50	5.05	329.9	31.1	38.7	5.6	40.3	9.6
Total	223.03	9.53	395.2	16.9	170.0	10.9	180.1	18.5
Ratio Estimator								
F	88.46	3.81	415.9	17.9	75.4	6.6	79.6	3.9
H	75.48	2.47	398.4	13.0	50.3	6.7	56.2	2.6
Other	48.36	4.52	298.2	27.9	34.4	5.6	34.7	4.4
Total	212.30	6.41	376.2	11.4	160.2	10.9	170.4	6.4
Regression Estimator								
F	91.84	3.08	431.7	14.4	79.9	5.0	83.3	3.6
H	78.78	2.60	415.9	13.7	51.7	5.8	59.2	2.5
Other	61.38	3.34	378.5	20.6	45.5	5.9	45.6	3.4
Total	231.99	5.26	411.1	9.3	175.7	9.6	187.3	5.2
Geometric Mean Regression Estimator								
F	91.65	3.06	430.8	14.4	79.4	3.2	82.7	3.0
H	79.32	2.74	418.7	14.5	54.7	3.3	58.7	2.7
Other	61.44	2.73	378.9	16.8	45.8	2.0	44.3	2.6

Total	232.44	4.93	411.9	8.7	179.9	5.0	185.8	4.9

[1] Total area of 564,306.4 ha with 212,713 ha in stratum F, 189,447 ha in stratum H, and 162,150 ha in stratum Other.
[2] The CWH Domain are the polygons identified as falling in the cedar-western-hemlock biogeoclimatic zone.

Because of the strong linear relationship between the Phase I and Phase II readings, it is not surprising that the precision of the ratio and regression estimators is improved compared to the simple inflation estimator. The regression estimator performs comparable to the ratio estimator as the fitted lines do tend to intersect the origin. There is little advantage to going to a geometric mean regression - the se are comparable to those from ordinary regression. As noted earlier, unless the error in X is considerable, there is little bias in ordinary regression. Domain estimates using Method 1 generally have a poorer precision than Method 2 - this is not unexpected and Method 1 should not be used if domain totals for the auxiliary variable are available (refer to Section C.15 for further details).

6. Discussion

Previous survey methods adopted the paradigm of "every hectare in the province should have an equal probability of selection". This lead to self-weighting estimators, but unfortunately, also made it very tempting to treat any survey obtained in such a fashion as being equivalent to a simple random sample of hectares when obtaining estimates and estimates of precision.

In this report, this requirement is relaxed in favor of an explicit multi-stage design whereby polygon are selected in the first stage with probability proportional to size with replacement (PPSWR), and ground points selected within polygons by any method that leads to unbiased estimates of polygon totals.

The advantages of this approach are:

(a) It gives greater weight to larger polygons as they have a greater impact upon the overall population parameters.

(b) It is flexible. Polygons can be added or removed from the survey design easily. The number of ground points sampled within polygons can vary from design

specifications without introducing additional problems. Methods used to sample ground points can vary among polygons. Allocation of number of sampling points among the sampled polygons can be varied to optimize overall precision requirements. Missing sampling points within polygons are easily accommodated as long at they occur completely at random.

(c) It allows estimates of precision to be easily computed regardless of complexities that may occur at the first or second sampling stage. These estimates of precision incorporate implicitly most sources of variation in the survey that occur.

(d) Computer software is available to assist in the analysis data collected from this design.

(e) Stratification of polygons by various attributes is explicit rather than implicit in previous methods. Allocation of sampling points among the strata is flexible to meet precision and other requirements rather than being fixed to ensure that every hectare receives and equal chance of selection.

It should be noted that the proposed methodology leads to algebraically identical estimators as proposed in previous reports when indeed every hectare has an equal probability of selection.

The potential disadvantages to this approach are, for the most part, minor:

(a) It is possible that polygons could be selected multiple times. However, this is not expected to occur very often given the large number of polygons in a typical inventory unit and the relatively small sample sizes selected.

(b) Computing polygon totals for variables such as age may seem odd - however, this was implicitly done in previous methods even though the simple formulae did not explicitly include polygon area or number of trees in the estimation formulae.

7. Further research

There are several areas in the VRI that were either glossed over lightly or not considered in this report that may require further research and analysis.

(1) Optimal allocation of number of sampling points within polygons. Previous survey methods always tried to give every hectare in the province an equal probability of selection. Consequently, larger polygons received more sampling points than smaller polygons. Because precision is dependent mostly upon absolute sample size rather than relative sample size, it is not necessary to maintain this ratio to obtain precise results. Previous surveys with multiple ground locations within polygon can be used to determine the optimal allocation of effort between sampling additional polygons and sampling ground points within polygons. Also refer to C.9

(2) Two-phase surveys whereby a sample of ground points is remeasured by another crew to determine the amount of measurement error have not been addressed in this report. The analysis and adjustment using this secondary survey can be complex was beyond the scope of this report.

(3) The report concentrated upon single variable ratio and regression estimators. Modern computer software now easily allows for multiple adjustment variables in complex survey designs. It may be advantageous to explore the use of additional adjustment variables.

(4) Estimated precision was derived using Taylor-series expansions. For small sample sizes these may not perform well. This report indicates that bootstrap methods are easy to implement because the polygons were selected with replacement. These and jackknife methods of variance estimation should be explored - particularly for small sample sizes that may be encountered in practice.

(5) This report did not examine the estimation or adjustment of categorical variables. Stauffer (1995, p. 51) also recommended that additional work be done on this problem.

(6) The problem of simultaneous adjustment of several variables for each polygon was also reviewed by Stauffer (1995, p. 38).

(7) Both this report and Stauffer (1995) did not consider the estimation and adjustment of compositional variables. For example, the stand composition of each polygon must add to 100%. The estimation, estimated precision, and adjustments for these types of variables is complex because of that additional constraint.

8. References

Angleton, G. M. and Bonham, C. D. (1995). Least squares regression vs. geometric mean regression for ecotoxicology studies. Applied Mathematics and Computation, 72, 21-32.

Berkson, J.B. (1950). Are there two regressions? American Statistical Association Journal 45: 164-180

Cochran, W. G. (1977), Sampling Techniques, Third Edition, New York: John Wiley & Sons, Inc.

Draper, N.R. and Smith, H. (1998). Applied regression analysis, 3rd edition. New York: John Wiley & Sons, Inc.

Kish, L. (1965). Survey Sampling. , New York: John Wiley & Sons, Inc.

Pathak, P.K.. (1962). On sampling units with unequal probabilities. Sankhya Ser A, 24, 315-326.

Penner, M. (2000) Procedures for handling unavailable sample sites in the Resources Inventory Program. Prepared for the British Columbia Ministry of Forests, Resource Inventory Branch, Victoria, BC.

Rao, J.N.K. (1997). Developments in sample survey theory: an appraisal. Canadian Journal of Statistics, 25, 1-21.

Riggs, D.S., Guarnieri, J.A. and Addelman, S. (1978). Fitting straight lines when both variables are subject to error. Life Sciences, 22, 1305-1360.

Sarndal, C.-E., Swensson, B., Wretman, J. (1992). Model assisted survey sampling. Springer-Verlag: New York.

Stauffer, H.B. (1995). The Statistical Estimation and Adjustment Process for the British Columbia Vegetation Resource Inventory. Prepared for the BC Ministry of Forests, Resources Inventory Branch.

Thrower, J.S and Associates (1998). Vegetation Resources Inventory Statistical Analysis: Final Report. Project MFI-401-038. Prepared for the Resources Inventory Branch, Ministry of Forests, 3 March 1998.

Appendix.A Glossary

Term	Definition
Accuracy	A measure of variation of an estimator around the true population value. Accuracy includes both sampling error and sampling biases. If an estimator is unbiased, then accuracy is the same as precision. If an estimator is biased, then it may not be accurate even if the precision is very good (i.e. has a small standard error).
Bias	Estimates never exactly equal the true (unknown) population value. Sometimes the estimate is larger than the population value; sometimes the estimate is smaller than the population value. If the average value of the estimate taken over all possible samples taken from the population equals the parameter, the estimator is said to be unbiased. If, on average, the estimates are smaller than the population value the estimator is said to be "negatively biased". If, on average, the estimates are larger than the population value, the estimator is said to be "positively biased". Bias is usually determined by theoretical means based on the sample design.
Confidence Interval	A range of plausible values for the true population value based upon information collected in the sample. A confidence interval has a level of confidence - by convention, 95% confidence intervals are found.
Domain	A sub-set of a population. For example, all polygons where significant insect damage has occurred could be a domain. Domains can be defined before the sample is selected and the population sub-divided into strata corresponding to the separate domains. Or, the domains can be defined after the sample is selected and domain estimation methods must be used.
Estimate	A statistic that is a "guess" for a population value. For example, the sample can be used to derive an estimated total volume of merchantable

	timber for an inventory unit. Every time a new sample is selected, the estimate will change (see sampling distribution).
Frame	A list of all the units in the population. For example, a list of all the polygons in the Inventory Unit. In multi-stage sampling, there will be one sampling frame per stage, i.e. a list of polygons in the inventory unit; a list of all the ground grid points within selected polygons, etc.
Parameter	A numerical value computed from the population units. For example, the total volume of merchantable timber in the inventory unit. Population parameters are always unknown and must be estimated from a sample.
Precision	Precision is a measure of sampling error, i.e. how variable are the estimates around their average value. It is commonly expressed by the standard error – an estimate with a smaller standard error is more precise.
Population	The set of all the units in the universe of interest. For example, all the polygons in the Inventory Unit. The population must be defined explicitly before a sample is taken, usually by the frame.
Random Sample	A method of selecting units for the sample in which every unit in the population has a chance of being selected which is known in advance. This term is often used (erroneously) as a synonym for a "simple random sample" in which every units has an equal chance of selection. As long as the probabilities of selection are known in advance, any sampling scheme that uses these probabilities of selection is a random sample.
Sample	The set of units selected for measurement. For example, the set of polygons selected from the Inventory Unit for which ground measurements will be taken.
Sampling distribution	Every time a new sample is taken, the statistics and estimates derived from the sample will change. The distribution of statistics or estimates over all possible samples is the sampling distribution of the statistic or

	estimator.
Standard error	The standard deviation of the sampling distribution. This is a measure of variability of an estimator around its average value. It measures sampling error only and does not include any bias effects.
Statistic	A numerical value computed from a sample. This is a generic term for any numerical derived from a sample and includes "estimates", i.e. all "estimates" are "statistics," but not all statistics are "estimates". Every time a new sample is selected, the statistics will change (see sampling distribution)
Stratum	A sub-set of the population. These can be defined before the sample is taken (pre-stratification) or after the sample is selected (post-stratification). If the case of pre-stratification, the surveyor has the ability to allocate samples among the strata to achieve pre-specified objectives. In the case of post-stratification, the sample sizes in each stratum are typically random. Refer also to "domains".

Appendix B– Notation

Symbol	Definition
a_h	The estimated intercept in stratum *h* for a regression estimator
b_h	The estimated slope in stratum *h* for a regression estimator
H	Number of strata in the population.
h	And index variable used to designate stratum. $h=1,...,H$
N_h	Number of units in the population in stratum *h*. For example, this would the number of polygons in each stratum.
n_h	Sample size in stratum *h*. For example, this would be the number of polygons selected and measured in each stratum.
p_i	Probability of selection of polygon *i* at each draw. This is found by the ratio of the polygon area to the total of the polygon areas for the stratum from which the polygon was selected.
R_h	The true (unknown) ratio between the Phase II and Phase I totals
$\hat{R}_h$	The estimated ratio between the Phase II and Phase I totals
X_i	Auxiliary variable for polygon *i* available from Phase I to be used in a ratio or regression estimator
X_h	Total of auxiliary variable in Stratum *h*
$\hat{Y}_i$	The estimated total for a variable in polygon *i*. This is obtained from the ground points surveyed. For example, this could be the estimated total volume of wood (m^3) for a polygon.
Y_i	The actual (unknown) total for a variable in polygon *i*. For example, this could be the actual (unknown) total volume of wood (m^3) for a

polygon. This is never known and can only be estimated.

Y_h	The true (unknown) total for polygons in stratum *h*
$\hat{Y}_h$	The estimated total for polygons in stratum *h*

Appendix C Frequently Asked Questions (FAQ)

C.1- Why can't I use the simple SRS formula for the estimates and standard errors?

> "Stauffer (1995, p.10) outlines estimators that are based on a simple random sample of grid points and states that such formulae give unbiased estimates and that the variance estimates should perform well."

Assume that polygons within stratum h were selected with probability proportional to area and that ground points within each polygon were selected using a simple random sample, i.e every point had an equal chance of being selected. Let y_i represent the measurement at ground location polygon i, a_i represent the area of polygon i and A_h represent the area of all polygons in stratum h. Then the probability of selecting a one-hectare unit on each sample:

$$\text{prob any hectare selected} = \frac{a_i}{A_h} \times \frac{1}{a_i} = \frac{1}{A_h}$$

i.e. every hectare in the inventory units stratum has the same probability of selection and is given an equal weight. Consequently, the estimated stratum total is computed by inflating the simple mean of all the data points:

$$\hat{Y}_{h,srs} = A_h \times \frac{\sum_{i=1}^{n_h} y_i}{n_h}$$

But, by rearranging terms, and noting that $a_i y_i = \hat{Y}_i$

$$\hat{Y}_{h,srs} = A_h \times \frac{\sum_{i=1}^{n_h} y_i}{n_h}$$

$$= \frac{1}{n_h} \sum_{i=1}^{n_h} A_h \frac{a_i y_i}{a_i}$$

$$= \frac{1}{n_h} \sum_{i=1}^{n_h} \frac{\hat{Y}_i}{a_i / A_h}$$

$$= \frac{1}{n_h} \sum_{i=1}^{n_h} \frac{\hat{Y}_i}{p_i} = \hat{Y}_{h,unf}$$

i..e. the two estimators are algebraically identical **assuming that the sampling plan proceeded without problems.** This was also noted by Stauffer (1995, p. 13). The advantage to expressing the estimator in th initial form is that there are **NO CHANGES** to the estimating equations even if every hectare is not given an equal chance of being selected. For example, under the equal-probability-for-each-hectare scheme, large polygons must have a larger number of ground points selected (on average) and no deviations within polygons are allowed, i.e. if a point cannot be sampled, great care must be exercised to choose another point within the same polygon. Under this more flexible scheme, the number of points sampled within each polygon can be chosen independently of the polygon size and there are no problems if the number of points actually sampled within each polygon differs from the theoretical specification. Consequently, there is no advantage to restricting the sample design so that every hectare has an equal chance of being selected given that the PPSWR design proposed is so flexible.

However, even though the two estimators may be algebraically identical when all assumptions for equal-hectare sampling are met, the theoretical precision is NOT computed as if each hectare were selected using a simple random sample because the actual sampling design is not a simple random sample, but rather is a two stage design. In practice, Stauffer (1995) argues "Each point on the ground has an equal probability of being sampled, and, although the sampling is not SRS... provide justification for this [the SRS] estimator [of precison]" on the ground that the previous method of first sorting polygons by categories and by polygon area and then using a systematic sample to select

ground points will lead to "the SRS estimator for the variance, though biased, will conservatively overestimate the variance". However, this argument will break down if the design does not provide that each hectare has an equal probability of selection, and, more to the point, why use an estimated variance that may be biased when the appropriate estimator for the variance is available and hardly more complicated to compute than the (invalid) SRS formula.

This same arguments can be used for the ratio, regression, and geometric mean regression estimators, i.e. (1) the given formula have an algebraically identical simpler form if every hectare has an equal probability of selection but the simpler form are not robust to violations of the assumptions, and (2) why use a formula for the estimated precision based on an inappropriate design when the correct formula for this design are readily available and are automatically computed by most computer packages?

C.2 What is the effect of fixed grid system?

> "Does the fixed provincial grid system cause a problem in either estimating the total or estimating the variance of the total?"

Sampling theory is based on the principle that random variation is always present and cannot be eliminated and so randomization ensures that, on average, the variation however caused, is represented in the sample in the same proportions as in the population. In the VRI, variation among ground values has been "aggregated" into a heirarchy – among polygons, among grid-points within a polygon, and measurement error. By randomly selecting polygons, randomly selecting points within polygons, and making sure that measurement errors are truly random and free of systemic bias, the estimators will be, on average, unbiased. Also, even though the formula for the variance of the estimators seems to "ignore" the latter two stages of variation, they have been accounted

for – what happens is that the variability captured by the lower stages is implicitly captured in the variation of the $\hat{Y}_i$ around the true polygon total.

However, the provincial grid defining the grid points is fixed and has not been randomized. Consequently, neither the estimator of the total, nor its estimated variance is guaranteed to be free of biases caused by the fixed grid system. Nevertheless, it is expected that such biases, if they exist, are small relative to other sources of bias. Similarly, the missing variation from the fixed grid system is also expected to be so small as to be effectively zero relative to the other sources of variation. There is of course no way of determining this from the current data – it would be possible to examine this assumption by taking some measurements off the current fixed grid and performing a components of variance analysis to actually measure the missing variation.

C.3 What is the sampling unit – a polygon or a point?

> "The actual measurements are taken at ground stations. Consequently, isn't the sampling unit a ground point and not a polygon"

A distinction should be made between the elemental units of the population and the way in which units are selected. The elemental units are the ultimate units of the population that are measured – in this case the ground points. The sampling units are elements or collections of elements from the population used to select samples. If samples were taken directly of the elemental units, i.e. if a list of every grid points in the province were available and a sample selected directly from this list, then the elemental units and the sampling unit would identical. However, in this protocol (as in any multi-stage design) there are several sizes of sampling units. At the first stage, polygons consisting of collections of ground points are selected – the sampling unit is a polygon. At the second stage, individual ground points are selected from the set within each polygon – the sampling unit is now the ground point.

If the ground point were a sampling unit, then the selection would have to be made at the first level at random from the entire set of all sampling points ignoring polygon boundaries.

Also refer to Section 2.4.7 for further details.

C.4 What happens if I can't sample all points within a polygon?

The current protocol and estimators are sufficiently flexible to accommodate changes in the number of points sampled within polgyons **as long as the estimate of the polygon total derived from the remaining points is unbiased.** Penner (2000) outlines several strategies to cope with missing sampling points – however, if there are multiple points within a polygon already measured and if the missing point is missing at random (i.e. unrelated to any attribute being measured) then the simplest solution may to simply drop the missing point and continue on.

Note that is a different problem than if a polygon is completely inaccessible, i.e. cannot be measured at any ground point. Refer to Section 2.4.5 for more details.

C.5 Why do I use only the estimated polygon totals and not the "point" value. Aren't I throwing away information?

> "The formulae in this report only use the estimated polygon totals. These could be based on 1 or 100 points within the polygon. This never seems to be used – isn't information being ignored?

As noted earlier in the report, the estimated polygon totals implicitly use all of the sampled ground points when used in the formulae for the overall totals. In particular, if each hectare in the stratum was selected with equal probability, then the simple random sample formulae are algebraically identical to those used in this report. In other cases the

simple random sample formulae are not appropriate and once the mean of the points is weighted by the polygon total, the formulae again reduce to those presented here.

What is very mysterious is the effect of the number of ground points. Presumably, sampling 100 points from every polygon leads to more precise estimates than only only sampling a single point from each polygon, yet the formulae for the estimated variance apparently fail to account for the number of points sampled from each polygon. In fact, the effects of different sample sizes are included implicitly.

Part of the confusion arises because of the distinction between the theoretical variance of an estimator and the estimated variance of an estimator. For example, in the most introductory statistics courses, student learn that the theoretical variance of the sample mean is $\frac{\sigma^2}{n}$ where σ^2 is the theoretical population variance of units in the population. Unfortunately, σ^2 is never known, and so the estimated variance of the sample mean is $\frac{s^2}{n}$ where s^2 is the sample variance. In the same way, the theoretical variance of the inflation estimator of the total in stratum *h* ignoring measurement error (from Cochran, 1977, p. 307):

$$V\left(\hat{Y}_{h,inf}\right) = \frac{1}{n_h}\sum_{i=1}^{N_h} p_i\left(\frac{Y_i}{p_i} - Y\right)^2 + \frac{1}{n_h}\sum_{i=1}^{N_h}\frac{a_i^2(1-f_{2i})S_{2i}^2}{m_i p_i}$$

where N_h is the number of polygons in stratum *h*; Y_i is the true total for polygon *i*; Y is the true population total; f_{2i} is the fraction of grid points sampled within polygon *i*; m_i is the total number of grid points sampled within polygon *i*; and S_{2i}^2 is the variability among grid points within polygon *i*. Now the effect of the number of grid points is clear – as m_i increases, the second term decreases and the theoretical variance decreases. Indeed, if m_i equaled all the grid points, the second term would vanish.

The estimated variance though is always of the form (Cochran, 1977, p.307):

$$v\left(\hat{Y}_{h,\inf}\right) = \frac{1}{n_h} \frac{\sum_{i=1}^{n_h} \left(\frac{\hat{Y}_i}{p_i} - \hat{Y}_{h,\inf} \right)^2}{n_h - 1}$$

So where has the second term in the theoretical variance gone? What happens is that as more and more ground points are sampled, the variation in $\hat{Y}_i$ declines (i.e if all grid points were sampled, then the exact value of the polygon total would be known), and the estimated variation would become smaller.

This is analogous to what happens in experimental designs with sub-sampling present – the analysis of variance table actually does not depend upon the sub-sample values and only the average over sub-samples is needed.

All of the above ignores measurement error. In theory, there would be a third term in the theoretical variance. It too is implicitly included as the variation in the $\hat{Y}_i$ includes all sources of variations below the polygon level.

The implications of these results is that decisions about the number of ground points measured per polygon do have impacts on the final precision. The number of ground points measured per polygon also have direct cost implications. In the current procedures, the number of ground points measured per polygon is fixed by the total sample size and the area of the polygon. However, the PPSWR design gives additional flexibility in the allocation of resources between the number of polygons selected and the number of ground points measured per polygon - this could be used to further improve precision for a given cost. Cochran (1977) has several sections on the allocation of resources between sampling at different stages of a multi-stage design.

C.6 Is it a problem that the aerial values are subject to "error"

> "Both the aerial photographs and the ground measurements measure the actual variable with error. Shouldn't a method that accounts for errors in both variables be used?

Design-based methods (such as the methods proposed in this report) remain unbiased regardless if both variables are subject to errors of measurement as long as there are no systemic biases. As well, Berkson (1950) showed that in certain circumstances, ordinary regression (where the X values are assumed to be known without error) still give consistent results. The use of aerial photographic readings to estimate the population total, is an example of a Berkson case, i.e. the aerial photographic readings (Phase I values) play the role of an X variable measured with error. All that happens is that the error of prediction is larger than necessary.

C.7 Isn't it better to stratify the population as much as possible?

Stratification is normally performed to increase precision of the overall estimate or because strata specific estimates are required.

In order for stratification to be successful in reducing the standard error of the overall estimate, the strata should consist of homogeneous polygons while the strata should be a different as possible. Consequently, there rapidly reaches a point of diminishing returns where after a few initial strata are defined, further strata are not that much different from each other. At this point, stratification should cease.

If stratification is being used to obtain stratum specific estimates, the number of strata can be as large as resources permit. However, if each stratum is to achieve a desired precision goal, the required sample sizes could be very large. As well, in order to determine an estimate of precision for each stratum, at least two polygons must be

sampled from each stratum. A large number of strata also makes for increased administrative complexity.

For example, consider the following table on the net volume per hectare stratified by leading species based on the Phase 1 readings from the SunShine Coast.

Summary Statistics on the net volume per hectare from the SunShine Coast			
Species	Polygons	Mean (unweighted)	Std dev (unweighted)
AC	289	145.1	112.3
AT	6	51.7	59.6
B	891	502.9	259.1
BA	969	72.5	188.9
BG	6	115.3	282.3
BL	3	239.4	124.4
CW	1679	322.0	264.6
DR	2857	230.7	132.3
EP	13	80.0	90.1
FD	10952	364.6	239.8
H	4313	447.2	219.9
HM	52	176.1	245.8
HW	4042	240.8	247.3
MB	209	272.9	128.5
PA	14	112.2	69.3
PL	551	130.9	103.7
PW	15	10.6	28.3
S	65	529.6	315.1
SE	36	8.5	50.7
SS	17	227.1	317.6
YC	239	330.6	249.5

The net volume per hectare varies considerably by leading species, but there is considerable variability within each stratum as well. Once the population has been stratified say by low, medium, and high densities, there is likely no benefit in terms of increased precision from further stratification. As each stratum needs at least two polygons to compute a variance estimator, it doesn't seem wise to allocate even that much effort to strata consisting of under 10 polygons whose contribution to the overall total is likely very small.

C.8 How many points should be sampled in each polygon?

Under the old protocols, great care was taken to give every hectare in the inventory unit an equal probability of selection. Consequently, the number of grid points in each polygon was proportional to its area, i.e polygons with twice the area received twice the grid points. However, under the proposed protocol, a great deal more flexibility is available.

The precision of each polygons estimated total depends primarily upon the total number of points sampled, not the relative fraction of points sampled. Consequently, there is no real advantage to measuring 10 points in very large polygons and only 5 points in smaller polygons. Indeed, given the homogeneity within polygons, there is likely little advantage to be gained from sampling more than one point in each polygon. Consider what would be the optimal strategy if the points within polygons were exactly identical. In this extreme case, there is no benefit from sampling additional points within a polygon – one point provides sufficient information to estimate the polygon total well. The "saved" effort would be more profitably put to use by sampling additional polygons.

If data existed on the within polygon variability from past studies, this "rule of thumb" above could be verified empirically.

C.9 What if the polygon boundaries change between the time the sample is selected and measured?

"The polygon boundaries changed between the time the sample of polygons was selected and the time the ground measurements were taken".

It may happen that because of the elapsed time between when the polygons are selected and the ground samples measured that the polygon boundaries could change. This is easily handled in the proposed protocol.

In these cases, the new polygons that include the sampled grid points should replace the old polygon and the sampling weights redefined to reflect the new polygons area.

C.10 Is it sensible to stratify by total volume as this is the primary attribute of interest?

> "Stratification leads to gains in precision when the polygons within strata are homogeneous and the strata are as different as possible. Consequently, are there any advantages to stratifying by total volume, i.e. one stratum could consist of large polygons with large total volume; a second stratum could consist of small polygons with very little total volume; etc"

Any stratification procedure that groups polygons into more homogeneous strata can leads to gains in precision. The degree of improvement is difficult to quantify as this depends upon the population value which are unknown. However, a reasonable approximation can be made by considering the Phase I data. For example, in the following table, three strata were created by sorting the Phase I total volumes in descending order by total volume and assigning the largest 499 values to the first stratum; the next 5000 to the second stratum; and the remainder to the last stratum. Then various allocations were examined in much the same way as in the detailed example of Section 5.

Effects of stratification by polygon area upon precision of the estimators

Stratum	N polygons	Total area	Variance	Equal Allocation n_h	Equal Allocation Var	Allocation Proportional to area n_h	Allocation Proportional to area Var	Optimal Allocation n_h	Optimal Allocation Var
A	499	44691.8	88.90	66.7	1.334	15.8	5.613	16.4	5.435
B	5000	170308.9	912.10	66.7	13.682	60.4	15.111	52.4	17.409
C	30345	349305.7	5724.00	66.7	85.860	123.8	46.236	131.3	43.611
	35844	564306.4		200.0		200.0		200.0	
Phase 1 volume		198.3							
		Theoretical Variance			100.875		66.959		66.455
		cv			5.1%		4.1%		4.1%
Non-stratification									
All	35844	564306.4	18363.7	200	91.818				
					4.8%				

This simple stratification could lead to a 30% reduction in the variance of the estimator which leads to a 15% reduction in the cv. As before, the gains from a proportional allocation to an optimal allocation are small. However, note that the sample size in the initial stratum may be insufficient if a further stratification is done by leading species.

Consequently, although in theory, this stratification can lead to gains in precision, there a practical problems that may render it moot. This stratification may be good for variables that are highly related to total volume – however, it may not perform well for other variables not related to total volume such as stand age or stand composition.

C.11 Why don't the estimators include stratum weights?

"Formulae for stratified random sampling found in text books include terms for stratum weights. Why don't these appear in the formulae in this report?"

The formulae presented in this report do include the stratum weights but implicitly because the totals are being estimated. For example, consider the standard formula for the estimated population mean from a stratified design with simple random sampling within each stratum:

$$\hat{\bar{Y}} = \sum_{h=1}^{H} W_h \bar{y}_h \text{ where } W_h = N_h / N$$

If, instead, the overall total was to be estimated, the overall mean would be multiplied by N giving:

$$\hat{Y} = N\hat{\bar{Y}} = N\sum_{h=1}^{H} W_h \bar{y}_h = \sum_{h=1}^{H} N_h \bar{y}_h = \sum_{h=1}^{H} \hat{Y}_h$$

because $N_h \bar{y}_h$ is an estimate of the total in stratum h. This latter formulae also appears to be lacking the stratum weights, but they implicitly included as well. A similar argument holds for the estimators in this report.

C.12 Why don't the variances formulae account for variability in polygon areas?

> "This report treats the polygon areas as fixed, known quantities. Yet polygon areas are subject to errors and variability - particularly when based on multiple layers in the GIS system. Why is this variability taken into account?"

First, it should be clarified that the differences in areas among polygons in the inventory unit is NOT a problem, i.e. the fact that there are small polygons and large polygons is not the "variation" that is of concern. Rather, a "theoretical polygon" may have an area of 10 ha, but, because of the way in which aerial interpreters draw polygon boundaries or other factors, it may be computed by the GIS system as 10.2 ha, or 9.5 ha etc. As long as there is no systemic bias, this should not cause any biases in the estimates. The variability caused by the variation in recorded polygon areas around the actual polygon areas IS included in the estimated precision for the same reasons as outlined in C.6. The theoretical variance formula will include terms for variation in polygon area around their

true values; the estimated variance formula implicitly includes this variation in the variation of the $\hat{Y}_i$. For example, if there were three identical polygons all of the same theoretical area, then the recorded areas of these polygons will differ and the $\hat{Y}_i$ will also reflect this additional source of variation.

C.13 How is choice made between the inflation, ratio, or regression estimator?

In any particular survey, many different estimates can be computed using different methods. How should the analyst choose among them? The choice among the inflation, ratio, and regression estimators should be made after an initial exploration of the data. To begin with, plots of the Phase I vs. Phase II variables will show if there appears to be any relationship between the two variables, if it is linear, and the approximate variance structure. If the relationship is very weak, then the inflation estimator will perform as well as the ratio or regression estimator - however, if the relationship is fairly strong, then the ratio or regression estimator should perform better than the inflation estimator and would be preferred. The choice among the regression and ratio estimator is less clear cut - unless there is very strong evidence that the intercept is not the origin, I suspect that both estimators will perform comparably. Other than these general comment, there is no objective criterion that is easily applied to indicate which is the preferred method.

C.14 What is the difference between a domain and a stratum or post-stratum

Strata are divisions of the entire inventory unit into sets of non-overlapping categories. The union of the strata gives the entire inventory unit. Strata can be determined before the sample is taken or after the sample is taken (post-stratification) - in both cases, the frame must contain sufficient information to allow each polygon in the unit to be allocated to its proper stratum. Strata are usually created because there is an intrinsic interest in the estimates for each stratum or as a variance reduction device.

A Domains is a subset of the entire inventory unit and estimates for this subset only are wanted.

For example, in this document, the inventory unit was stratified by leading species (F, H, or other). The domain of interest were polygons belonging to the cedar-western-hemlock biogeoclimatic zone. Note that a domain can cut across strata.

There is a close relationship between strata and domains. One could think of a domain as one of two strata - a polygon is either a member of the domain or not a member of the domain. However, domains are usually defined after the sample is taken and are 'ad hoc' groupings. Consequently two methods of domain estimation are often required in case the information about domain membership for polygons is or is not available for every polygon in the unit.

C.15 When should Method 1 and Method 2 be used for domain estimation?

The choice between the two methods is primarily based on the availability of sufficient information about the domain at the frame level. Method 2 requires that every polygon in the frame be classified as to domain membership so that the number of polygons and the total area within the domain is known. Method 1 only requires total for the entire inventory unit.

For example, if a domain is defined by the amount of insect damage and it is only possible to determine this from ground measurements, then Method 2 cannot be used.

If sufficient information is available, then Method 2 is preferable as it usually gives more precise estimates.